Title

Remarkably Divergent Regions Punctuate the Genome Assembly of the Caenorhabditis elegans Hawaiian Strain CB4856.

Journal

Genetics, 200(3)

ISSN

0016-6731

Authors

Thompson, Owen A
Snoek, L Basten
Nijveen, Harm
et al.

Publication Date

2015-07-01

DOI

10.1534/genetics.115.175950

Peer reviewed

Remarkably Divergent Regions Punctuate the Genome Assembly of the *Caenorhabditis elegans* Hawaiian Strain CB4856

Owen A. Thompson,* L. Basten Snoek,† Harm Nijveen,‡ Mark G. Sterken,† Rita J. M. Volkers,† Rachel Brenchley,§ Arjen van't Hof,§ Roel P. J. Bevers, Andrew R. Cossins,§ Itai Yanai,†† Alex Hajnal,‡‡ Tobias Schmid,‡‡ Jaryn D. Perkins,§§ David Spencer,* Leonid Kruglyak,*** Erik C. Andersen,††† Donald G. Moerman,§§ LaDeana W. Hillier,* Jan E. Kammenga,† and Robert H. Waterston*,1**

*Department of Genome Sciences, University of Washington, Seattle, Washington 98195, †Laboratory of Nematology, Wageningen University, 6708 PB Wageningen, The Netherlands, ‡Laboratory of Bioinformatics, Wageningen University, NL-6708 PB Wageningen, The Netherlands, §Centre for Genome Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, United Kingdom, **Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland, ††Department of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel, ‡‡Institute of Molecular Life Sciences, University of Zurich, CH-8057 Zurich, Switzerland, §§Department of Zoology and Michael Smith Laboratories, University of British Columbia, Vancouver, BC, Canada V6T 1Z3, ***Howard Hughes Medical Institute, Department of Human Genetics and Department of Biological Chemistry, David Geffen School of Medicine, University of California, Los Angeles, California 90095, and †††Department of Molecular Biosciences, Northwestern University, Evanston, Illinois 60208

ABSTRACT The Hawaiian strain (CB4856) of *Caenorhabditis elegans* is one of the most divergent from the canonical laboratory strain N2 and has been widely used in developmental, population, and evolutionary studies. To enhance the utility of the strain, we have generated a draft sequence of the CB4856 genome, exploiting a variety of resources and strategies. When compared against the N2 reference, the CB4856 genome has 327,050 single nucleotide variants (SNVs) and 79,529 insertion–deletion events that result in a total of 3.3 Mb of N2 sequence missing from CB4856 and 1.4 Mb of sequence present in CB4856 but not present in N2. As previously reported, the density of SNVs varies along the chromosomes, with the arms of chromosomes showing greater average variation than the centers. In addition, we find 61 regions totaling 2.8 Mb, distributed across all six chromosomes, which have a greatly elevated SNV density, ranging from 2 to 16% SNVs. A survey of other wild isolates show that the two alternative haplotypes for each region are widely distributed, suggesting they have been maintained by balancing selection over long evolutionary times. These divergent regions contain an abundance of genes from large rapidly evolving families encoding F-box, MATH, BATH, seven-transmembrane G-coupled receptors, and nuclear hormone receptors, suggesting that they provide selective advantages in natural environments. The draft sequence makes available a comprehensive catalog of sequence differences between the CB4856 and N2 strains that will facilitate the molecular dissection of their phenotypic differences. Our work also emphasizes the importance of going beyond simple alignment of reads to a reference genome when assessing differences between genomes.

KEYWORDS *C. elegans*; genome assembly; evolution; variation

DNA sequence variation, whether present in natural populations or induced in the laboratory, has been central to the functional understanding of genes and genomes. Natural variation has proven particularly valuable in the analysis of quantitative traits while also providing insights into the evolutionary processes that shape genomes. At the same time, mutations of strong phenotypic effect have long been a pillar of experimental genetics. As rapidly improving DNA sequencing technology has simplified both the detection and the cataloging of variation, major efforts have been undertaken to describe variation and then analyze quantitative traits in wild isolates of various model organisms, including *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Drosophila*, and *Arabidopsis* (Schacherer *et al.* 2009; Cao

doi: 10.1534/genetics.115.175950
Manuscript received March 2, 2015; accepted for publication April 29, 2015; published Early Online May 19, 2015.
Available freely online through the author-supported open access option.
Supporting information is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1.
Sequence data from this article have been deposited in GenBank under accession no. JZEW00000000 (CB4856 genomic assembly) and in Sequence Read Archive under accession no. SRX1001806.
[1]Corresponding author: Department of Genome Sciences, University of Washington School of Medicine, 3720 15th Ave. NE, Box 355065, Seattle, WA 98195-5065. E-mail: watersto@u.washington.edu

et al. 2011; Andersen *et al.* 2012; Mackay *et al.* 2012; http://www.1001genomes.org).

For *C. elegans*, the canonical wild-type laboratory strain N2 was derived from an isolate found in 1951 in mushroom compost in England (Nicholas *et al.* 1959) and maintained in liquid culture on agar slants and then on *Escherichia coli* until protocols were developed in 1969 that allowed storage of frozen stocks (Sulston and Brenner 1974; Sterken *et al.* 2015). It was the first multicellular organism to have a fully sequenced genome (*C. elegans* Sequencing Consortium 1998), and this sequence has served as the reference for *C. elegans*. Prior studies that have utilized the genetic diversity available in wild populations of *C. elegans*, whether studying specific phenotypes, genes, gene classes, or population mutational spectra, have reported their results with respect to the N2 strain.

Among the many wild isolates of *C. elegans*, one of the most genetically divergent and most heavily studied is CB4856, which was isolated in 1972 by Linda Holden from a pineapple field on the Hawaiian island of Maui (under the name HA8) (Hodgkin and Doniach 1997). It shows multiple phenotypic differences with N2, including production of a copulatory plug, aggregation behavior, a lack of temperature-size dependence, growth rate, fecundity, RNA interference insensitivity by feeding and drug resistance (de Bono and Bargmann 1998; Kammenga *et al.* 2007; Ghosh *et al.* 2012; Pollard and Rockman 2013; Andersen *et al.* 2014), and gene expression differences (Capra *et al.* 2008; Rockman *et al.* 2010; Vinuela *et al.* 2012; Volkers *et al.* 2013). Various populations of recombinant inbred lines (RILs) and a population of introgression lines (ILs) have been generated between CB4856 and N2 to define the genetic architectures of complex genetic traits (Li *et al.* 2006; Rockman and Kruglyak 2008; Doroszuk *et al.* 2009; Andersen *et al.* 2015). Molecular genetic analyses of the Hawaiian strain have revealed polymorphisms associated with several of the above traits as well as others. An online database, WormQTL, has been created for the deposition of expression quantitative trait loci (Snoek *et al.* 2013, 2014; van der Velde *et al.* 2014).

The elucidation of sequence variants in CB4856 has occurred in several steps. Initially, random genomic fragments were compared to the N2 reference genome, revealing >6000 SNVs and small insertion/deletions (indels) (Wicks *et al.* 2001). A later study increased the number of SNVs to >17,000 (Swan *et al.* 2002). The genomic positions of these SNVs are distributed nonrandomly, with more variation present on chromosome arms than in the centers where recombination is lower (Koch *et al.* 2000; Wicks *et al.* 2001). These variants provided suitable markers for genetic mapping using a variety of methods. D. Spencer and R. H. Waterston (unpublished results) cataloged >100,000 SNVs using an early version of massively parallel sequencing (MPS) technology in a whole-genome shotgun (WGS) approach and deposited these variants in WormBase, noting multiple ~25- to 100-kb regions of poor read alignment, possibly due to high sequence divergence. These regions were most prevalent on the left arms of chromosomes I and II along with both arms of chromosome V. Array comparative hybridization identified large copy number variations (CNVs) and found that these CNVs also were enriched on chromosome arms, affecting primarily gene family members that had undergone recent expansion in *C. elegans* (Maydan *et al.* 2007, 2010). A study of chemoreceptor gene families uncovered functional genes in CB4856 that are defective in N2 (Stewart *et al.* 2005). Recent genomic analyses of CB4856 and N2, alongside other isolates, again found the Hawaiian strain to be among the most divergent, either by using sequencing restriction-site-associated DNA markers in 202 strains (Andersen *et al.* 2012) and/or by comparing hybridization of coding sequences between N2, CB4856, and a panel of 46 wild isolates (Volkers *et al.* 2013). Recently, we used MPS to obtain deep WGS coverage, providing a more complete list of differences including indels of a full range of sizes between the N2 reference and the Hawaiian genome (175,097 SNVs and 46,544 indels) (Thompson *et al.* 2013). Another group extended the set further using deeper WGS coverage along with longer reads from the 454 platform (Vergara *et al.* 2014).

One shortcoming of all of these studies has been that they have relied on alignment of the sequence reads to the N2 reference genome. As a result, multiple regions of the Hawaiian genome remain missing or poorly defined. These missing regions include insertions in the Hawaiian genome relative to N2. But, in addition, inspection of the deep WGS coverage revealed some regions of the genome that apparently were so divergent that aligned reads were sparse to absent (D. Spencer, O. A. Thompson, and R. H. Waterston, unpublished results). The sequence of these highly divergent regions and Hawaiian specific sequences must be determined to interpret more fully any genotypic and phenotypic differences between the Hawaiian and N2 strains.

Accordingly, we have undertaken the construction of a Hawaiian reference genome sequence that more completely reflects the sequence differences between the two isolates. To accomplish this goal, we took advantage of several very deep coverage MPS data sets for the Hawaiian genome, a *de novo* assembly program (Chu *et al.* 2013), end sequences from a fosmid library for the Hawaiian genome, recently released RNA-seq data, and low-coverage genome sequence data from 49 RILs (Li *et al.* 2006) and 60 ILs (Doroszuk *et al.* 2009) (Table 1). Exploiting these resources and using a variety of software tools, we have modified the N2 reference genome to generate a draft reference sequence for the Hawaiian genome. The results reveal >60 regions with haplotypes that are substantially divergent from N2. The distribution of these haplotypes in other wild isolates suggests that these regions were present in the genomes of ancestral populations before the world-wide distribution of the *C. elegans* species and have been maintained since that time.

Table 1 CB4856 sequence resources

Data set	PI	Type	Platform	[SlP]E, length (bp)	Insert size	Clones/total bases in reads	Coverage expected (%)
Princeton University	Andersen	DNA	Illumina	PE 104, 104	321 bp	34,711,778/7,220,049,824	69.52× (96.6)
University of Washington (Thompson *et al.* 2013)	Waterston	DNA	Illumina	PE 76, 76	179 bp	21,252,827/3,230,429,704	31.10× (96.6)
Technion	Yanai	DNA	Illumina	PE 100, 100	221 bp	79,406,930/15,881,386,000	80.08× (50.6)
University of Zurich	Hajnal	DNA	Illumina	PE 101,101	484 bp	825,754/166,799,884	1.41× (84.9)
University of Zurich	Hajnal	DNA	SOLiD	PE 50, 35	124 bp	15,760,405/2,679,268,850	7.20× (26.9)
University of British Columbia (Perkins 2010)	Moerman	DNA	Sanger	PE ~770 bp	~33 kbp	15,360/20,520,434	0.20× (97.2)
Washington University (Wicks *et al.* 2001)	Waterston	DNA	Sanger	SE ~764 bp	NA	11,541/8,843,526	0.07× (81.7)
Wageningen University/ University of Liverpool	Kammenga/ Cossins	DNA (ILs/RILs)	SOLiD	SE 50	NA	2,709,932,329/135,496,616,450	766.85× (56.8)
Total DNA							956.43×

SE = single end
PE = paired end

Materials and Methods

Sequencing methods

DNA from the *C. elegans* CB4856 strain was extracted using the Qiagen Blood and Tissue kit and quantified using a Qubit 2.0 broad-range kit. DNA was sheared in a Covaris LE220 sonicator to a size of 300–600 bp and then 400- to 600-bp fragments were gel-extracted after standard Illumina TruSeq sample preparation. One Illumina HiSeq2000 lane was run to obtain the 101-bp paired-end sequence reads used in this study.

Creating a Hawaiian reference

Using a strategy similar to that employed in the analysis of different *Arabidopsis* accessions (Gan *et al.* 2011; Schneeberger *et al.* 2011), we first aligned the random genomic reads to the N2 reference genome, identified SNVs and indels, modified the N2 reference accordingly, and realigned the reads, repeating the process 19 times to create a first version of the Hawaiian genome (20 cycles total) (Figure 1; Supporting Information, Figure S1; Table S1). This process allowed extension of sequence into regions of high divergence, closed large deletions, and built sequence into insertions (Figure 1 and Figure 2). We used the JR-Assembler (Chu *et al.* 2013) to create *de novo* assemblies of the same sequence reads, assessed their quality using the program REAPR (Hunt *et al.* 2013), breaking contigs as needed, and aligned the resultant contigs to the Hawaiian genome. To identify deletions previously missed, we scanned the genome for regions devoid of coverage, merging adjacent regions if they were separated only by short segments of either very low coverage or repeated sequences. For regions flanked by adjacent segments of the *de novo* assembled contigs, we used the contig to close the gap. To confirm that such segments were properly placed in the genome, we used the RIL data to establish their chromosomal location (Figure S2; Figure S3; Table S2). Specific methods for each of these steps are presented in the Supporting Information.

The result is an initial draft reference Hawaiian genome with a total length of 98.2 Mb. Regions of excess coverage (>99×) suggest that we have failed to represent some duplicated segments, which total some 0.5 Mb in length. Also, the *de novo* assembly generated 22 contigs of 16 kb total length that we were unable to locate in the reference. Just as the N2 reference has been improved through continuous community input, we expect users will provide improvements here.

ALE scoring of divergent regions in wild isolates

MPS sequence reads from each of the 39 wild isolates previously studied (Thompson *et al.* 2013) were aligned against both the N2 and CB4856 reference sequences. The resultant alignments were scored using the ALE program (Clark *et al.* 2013). For each divergent region (Table S3), we then plotted the placement score for each strain against the two genomes. Many sites followed a simple binary pattern, with scores for each strain against N2 and CB4856 resembling one of the controls (Figure S4A). Other regions in some strains showed intermediate scores against either or both N2 and CB4856. Inspection of the alignments and ALE score patterns across the region suggested that the strain had intermediate divergence across the region, where some blocks of a region resembled the N2 haplotype and others resembled CB4856 (Figure S4B). These patterns were consistent with the idea that recombination had occurred within the region. However, some regions had one or more strains with reads that aligned poorly with both strains, and inspection of those regions in these strains was consistent with the presence of a third haplotype for the region (Figure S4C). Other regions, particularly those from the left arm of chromosome II, had more complex patterns and were not analyzed further.

Comparing the two reference genomes

The *C. elegans* N2 genome, version WS230 (note that while annotations changed, the N2 genome sequence remained

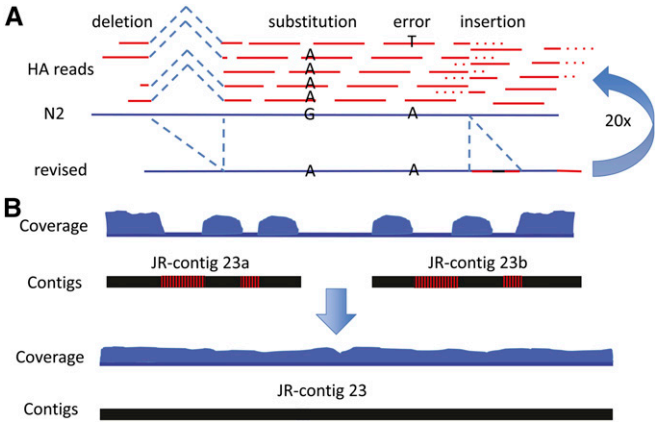


Figure 1 Strategy for constructing a Hawaiian reference sequence. (A) Alignment of 100-bp paired-end reads from the CB4856 genome to the N2 genome. Sites that differed by base substitution and insertion and deletion were recognized, and the N2 genome was altered at those sites. For insertions larger than a read and at the edge of divergent regions, the consensus sequences from the unmatched segments of the reads were added to the reference. Then the reads were aligned to the modified reference, and the cycle was repeated for 20 times, by which time few changes were being made. (B) After the 20 cycles of alterations, areas with incomplete coverage still persisted. To correct these areas, individual reads were assembled *de novo* with the JR-Assembler and aligned against the modified reference. Typically, these JR contigs would show good agreement where read coverage was good, and thus corrections had been made, but poor alignment where the reference sequence did not have coverage and had not been altered from the N2 reference. The JR contigs were also aligned against sequence reads from RILs and ILs. Only RILs and ILs containing a segment of the Hawaiian genome that spanned the JR contig yielded good coverage across these divergent regions, thereby locating the JR contigs on the genome. Where the JR contigs had regions of good match against the reference and their location was confirmed by alignment of reads from RILs and ILs, they were spliced cleanly into the reference. Remaining large deletions were also removed.

identical from WS215 through WS234), was aligned against the Hawaiian genome using LASTZ [version 1.02.00; O = 400, E = 30, K = 3000, L = 3000, M = 0; (Harris 2007)]. The program LASTZ, like BLASTZ (Schwartz *et al.* 2003), uses a series of steps consisting of seeding (a user-specific seed pattern is allowed), gap-free extension, chaining, anchoring, gapped extension, and interpolation. However, unlike BLASTZ, LASTZ can derive its own suitable tuning parameters from the sequences themselves. We chose LASTZ because it is able to perform pairwise gapped full chromosome-to-chromosome alignments using very little memory.

The LASTZ alignments were performed with the chain option and lav output format. Files in the lav output format were converted to psl format using lav2psl (J. Kent, http://hgdownload.cse.ucsc.edu/admin/jksrc.zip) and were chained [creating a sequence of gapless aligned blocks with no overlapping blocks (Kent *et al.* 2003)] using axtChain (-linearGap = loose). The resulting alignment files (a readable text version for coordinate lookup is available in File S1) were merged using chainMergeSort and prepared for the netting step using chainPreNet and then netted (a net is a hierarchical collection of chains, with the highest-scoring nonoverlapping chains on top; their gaps were filled in where possible by lower-scoring chains; for more information see http://genomewiki.ucsc.edu/index.php/Chains_Nets), retaining a unique, ordered set of alignments within the N2 genome using chainNet. The resulting output contains a single set of ordered alignments in the N2 genome followed by additional "fill in" alignments. Only alignments to the same chromosome in N2 and Hawaiian and on the same strand were retained. We reviewed the additional "fill in" alignments and found only a possible 2 kb of unique additional aligned sequence that could have been used for SNV identification. Therefore, we chose not to retain the "fill in" alignments (they occur in the output file after the initial list of the entire N2 chromosomal alignments). The axt files created by the final netting step were parsed using a custom script to create files listing the SNVs and insertion/deletion differences.

When there were gaps in either the N2 or Hawaiian genome between the "major" alignments, they were annotated as indels. Two types of regions were identified between neighboring blocks: (1) simple indels, where the sequence was cleanly inserted/deleted in one genome relative to the other and (2) 816 cases where sequence was present in both the N2 and Hawaiian genomes but could not be accurately aligned (alignment using various Smith–Waterman algorithm implementations revealed different alignments in each case). Thus, we elected to simply retain those regions as blocks of indels. To identify high-quality substitutions within the LASTZ alignments, we required at least three reads and ≤150 reads with the fraction of reads that disagree with the total reads at the site as <0.2.

RNA-seq ALIGNMENTS/TOPHAT/CUFFLINKS

Hawaiian reads from available RNA-seq projects (313,124,440 reads) (Stoeckius *et al.* 2014) were aligned to the Hawaiian genome using TopHat (–b2-fast) (Trapnell *et al.* 2012), and BAM files were generated containing the resulting 215,316,914 aligned reads. Cufflinks (–min-frags-per-transfrag 5–min-intron-length 25–trim-3-avgcov-thresh 5 -p 8) created 31,965 transcript predictions (30,199 genes) composed of 91,185 exons (84,461 different exons) (Trapnell *et al.* 2012).

Genefinder

Using the *C. elegans* table files, we ran Genefinder (version 1.1; P. Green, unpublished results; -orfcutoff -0.5 -intron3-cutoff -0.5 -intron5cutoff -0.5) to provide *de novo* gene prediction in the Hawaiian genome. A total of 22,261 transcripts were predicted, which were composed of 126,518 different exons.

Identifying N2 genes in Hawaiian

To identify *C. elegans* N2 genes in Hawaiian, the N2 predicted genes (WS240) were aligned against the Hawaiian genome using blat (-q = rna). Each N2 gene was aligned to

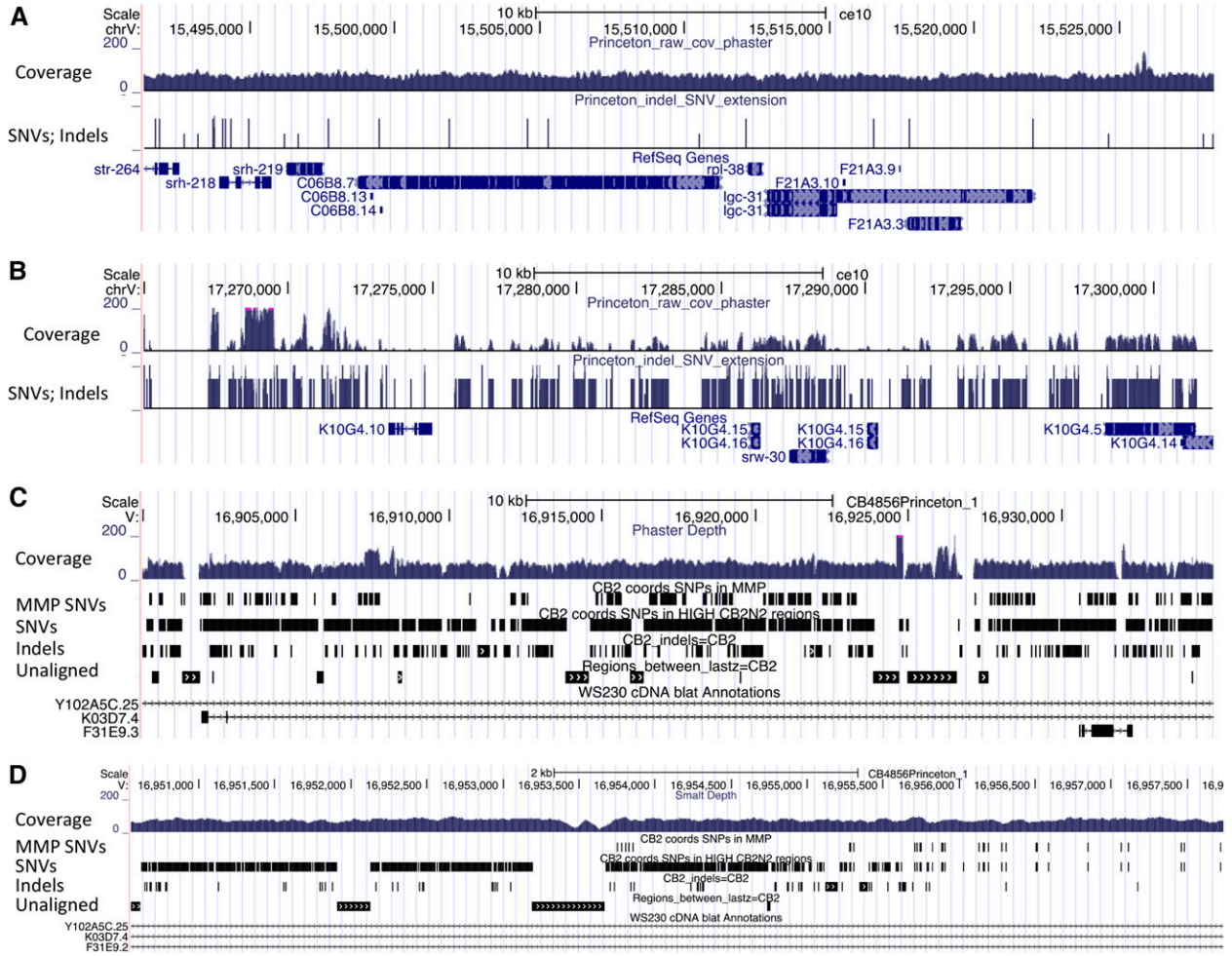


Figure 2 Read coverage and SNV density in the N2 reference genome and the iteratively corrected CB4856 genome. (A) A typical region for most of the genome is shown, with good coverage (top track) and infrequent SNVs and indels (second track). Genes are shown below. (B) A region of the N2 reference showing poor coverage and a high SNV/indel density with the Hawaiian reads. (C) After 20 iterations of reference-guided corrections, the same region as in B now has improved coverage by the CB4856 reads. In addition to coverage, the tracks show the SNV calls (MMP SNVs) reported in (Thompson *et al.* 2013), the SNV calls based on the new reference (SNVs), indels based on the new reference (Indels), and regions that failed to align with sequence present in the N2 reference (Unaligned). Gene models for each region are shown below. (D) The boundary of a divergent region (left) with a less divergent region of the genome is shown. The density of SNVs and indels changes abruptly. Tracks are as in C.

the corresponding Hawaiian chromosome. At least partial alignments were obtained for 26,571 of the 26,769 genes.

Validation of CB4856 by random long (Sanger) reads—SNVs

The CB4856 Sanger sequencing reads from the Washington University (random genomic short inserts) and the University of British Columbia (insert end sequences of a genomic fosmid library) projects were aligned against the final CB4856 reference using bwa (Li and Durbin 2009; Li *et al.* 2009). To assess whether the Hawaiian SNVs presented here were confirmed by those long reads, all reads with a mapping quality of >40 (25,667 total reads with 13,835,952 bases having phred quality scores ≥40) were retained for analysis. A total of 28,203 of the Hawaiian SNVs presented here were covered with at least one of the base calls from the long read set of phred quality ≥40. Of those 28,203, there were 27,680 (98.1%) where the long read confirmed the Hawaiian SNV. Of those, 8380 were in a divergent region, and 8123 of those were confirmed (96.9%). For the 19,823 not in divergent regions, 19,556 were confirmed (98.7%). For the 523 positions where the Hawaiian SNV was not confirmed by the long read data, we examined the basecall from the JR-Assembler for that position in the genome. Of the 523 positions, there were 416 total (172 in the divergent regions) where the position was in a JR-Assembler contig that was not one of the 221 complete JR-assembled contigs inserted into the final reference sequence. Small parts of other contigs were also incorporated into the final assembly, but because they were small, they were not excluded here. In 415 of the 416 cases, or less than half a percent of the cases, the base called in the JR-Assembler contig agreed with the Hawaiian SNV call. The single case that did not agree was in one of the divergent regions. The consistency between the

integrated assembly and the JR-contigs suggests that many of the differences with the fosmid end sequences could be due to strain differences or from errors in the Sanger reads.

Validation of CB4856 by random long (Sanger) reads—indels

There are 190 reciprocal indels where a bwa alignment suggests a long Sanger sequencing read alignment spanning a position of a reciprocal deletion (106 of which are in divergent regions). In each case, the reads involved were realigned with the CB4856 genome using cross_match (P. Green, unpublished data), providing an alternative alignment method and a full Smith–Waterman alignment of the read against the genome, and analyzed. Of those 190, in 3 (1.6%) cases (1 of the 106 in the highly divergent regions) the alignment of the long read suggests an alternate indel from what is in the CB4856 assembly.

There are 4900 simple insertions (1196 in highly divergent regions) where a bwa alignment identified a read that aligns within the simple insertion. After alignment with cross_match, in 11 cases (1%) the long read suggests an alternate insertion [4 (0.3%) in highly divergent regions].

Results

Generating a Hawaiian reference sequence

To develop a reference sequence for the CB4856 genome that would account for regions of poor representation and incorporate sequence present in CB4856 and missing in N2, we exploited data sets from across the world (Table 1). The missing and poorly defined regions of the CB4856 genome might reflect problems with specific libraries or sequencing platforms. Aligning sequence reads against the N2 reference genome from three independently generated libraries sequenced with the Illumina platform gave broadly similar patterns of coverage, with similar regions of poor or no coverage of the N2 reference. Calling SNVs on these data sets yielded essentially the same set of SNVs with no more than 3% of SNV calls unique to any one library (Figure S5). These regions of poor coverage persisted with different aligners that handle repeats differently—*phaster*, which places all reads that match equally well at multiple copies of identical repeats at the first copy (P. Green, personal communication), and bwa or *smalt* (http://sanger.ac.uk/resources/software/smalt/), which distributes the matching reads randomly among the copies. The small percentage of SNVs that were unique to individual data sets tended to be in regions where read coverage was marginal (*e.g.*, Figure 2), leading to above-threshold calls in one data set, but not in the others. Similar regions of poor coverage were also seen with the WGS library sequenced with SOLiD technology. These regions were also flagged with tools designed to detect alignment irregularities, such as REAPR and ALE (Clark *et al.* 2013; Hunt *et al.* 2013). Because these regions of poor coverage persisted across libraries and platforms, we postulated that these regions might reflect segments of the genome with much higher-than-average sequence differences that led to variability in read alignment or even alignment failure.

Table 2 Comparison of reference sequence lengths

Chromosome	N2	HA	Difference	%
I	15,072,423	14,890,789	181,634	1.21
II	15,279,345	14,885,952	393,393	2.57
III	13,783,700	13,596,826	186,874	1.36
IV	17,493,793	17,183,857	309,936	1.77
V	20,924,149	20,182,852	741,297	3.54
X	17,718,866	17,537,347	181,519	1.02
Total	100,272,276	98,277,623	1,994,653	1.99

% difference in size expressed as a percentage of the length of the chromosome in the Hawaiian genome.

To improve read alignment in such regions and to extend coverage, we initially employed a technique similar to the reference-guided assembly strategy used in *Arabidopsis* (Gan *et al.* 2011; Schneeberger *et al.* 2011). This guided assembly approach allowed us to exploit the continuity and high quality of the N2 reference sequence and to avoid the pitfalls associated with whole-genome assemblies using current technologies and algorithms. We used the called SNVs to change the N2 reference sequence to reflect the presumptive Hawaiian sequence throughout the genome (Figure 1). We also deleted N2 sequence where read coverage was largely lacking and split reads clearly spanned from one N2 region to another. We added sequence at the edge of insertion sites, based on the consensus sequences of the unmatched portions of reads. We then realigned the reads against the revised sequence. For these purposes, we used only a single data set to avoid possible differences between starting strains and to avoid the excessive computational demands from using all the WGS data sets. We selected the Princeton data set because it had very deep coverage (~70×), longer paired reads with larger inserts, and a high fraction of reads aligning to the genome. An initial cycle significantly improved alignments, and after 20 cycles of alignment and correction most regions had excellent agreement between the draft reference and the sequence reads. For example, alignment of CB4856 reads against the draft reference produced only 13 SNVs and 8 indels (compared with 219,787 SNVs and 46,674 indels against the N2 reference). As a broader measure of the improvement, average ALE scores per base, adjusted using N2 reads against the N2 reference as a baseline, decreased from −6.10 to −2.49 (Figure S6).

Even after 20 cycles of replacements, we found ~50–70 regions of 10–100 kb where the initial uneven coverage of the N2 sequence had nucleated improved coverage after the iterative correction/extension approach, but overall coverage of the region remained discontinuous; ALE and REAPR scores remained high in these regions (Figure 2). These regions often had indications of multiple deletions that lacked spanning reads to define their end points. As a result, they had not been removed in our iterative alignment and

Table 3 Number of deletion events and base counts in deletions in N2 and CB4856

	Deletions in N2		Deletions in CB4856	
Chromosome	Events	Bases	Events	Bases
I	5,693	94,543	6,158	233,930
II	7,370	230,813	7,692	478,884
III	5,464	99,324	6,008	275,997
IV	5,700	116,249	5,841	265,504
V	10,902	343,640	11,740	854,733
X	3,453	37,442	3,507	156,819
Total	38,582	922,011	40,946	226,5867

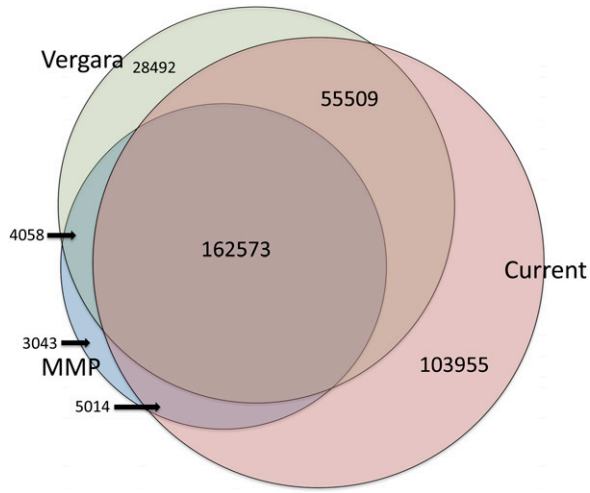


Figure 3 Overlap with previous SNV calls. A Venn diagram shows the overlap of the previous SNV calls with those obtained with the CB4856 reference.

correction procedure. To resolve these regions, we exploited a new whole-genome assembly algorithm, the JR-Assembler (Chu *et al.* 2013), which compares favorably to current assemblers in terms of quality, efficiency, and memory. We used the JR-Assembler to create a *de novo* assembly of the CB4856 genome. To reduce the number of likely false joins, we analyzed the assembled JR contigs with REAPR, splitting them where REAPR signaled problems.

To utilize the JR contigs, we first aligned the resultant 14,167 JR contigs, totaling 93,075,504 bases with an N50 of 15,785 bp (1601 contigs) against both the N2 and the iteratively improved genomes. The JR contigs aligned against N2 showed base-pair disagreements consistent with the called SNVs in most regions of the genome, *i.e.*, those that had good coverage with the CB4856 sequence reads. However, they revealed additional SNVs and indels in the 50–70 regions of spotty coverage (Figure 2). In contrast, when the JR contigs were aligned against the recursively corrected genome, they showed only rare differences, except in the regions with spotty coverage. In those regions, the JR contigs indicated additional base-pair changes and likely indels, suggesting that substituting the JR-contig sequences in the divergent regions of the recursive genome could be used (1) to remove some deletions that had failed to meet our criteria for removal in the iterative process, (2) to add sequence that had not been fully added in the 20 cycles, and (3) to correct some remaining substitutions (Figure 1 and Figure 2).

To confirm that the JR contigs were properly placed within the genome, we examined coverage metrics for each contig across a collection of 60 introgression lines and 49 RILs. For contigs aligning to these divergent regions, normal coverage with Hawaiian-derived reads was limited to those strains with a Hawaiian segment for that region, thus locating the contig within a region of a few megabases (Figure S7). Similar metrics were used to confirm placement of large insertions. Using only confirmed contigs, we replaced sequences in the recursive genome with the sequences from the JR contigs totaling >2 Mb (see *Materials and Methods* for details). We also removed presumptive deleted regions that had no JR contigs aligning and no aligning reads. These changes dropped the adjusted genome-wide ALE scores from −2.48 to −1.18 (Figure S6), and visual inspection of the divergent regions also supported a much-improved version of the genome (illustrated in Figure 1). Alignment of 42,261 Sanger reads from fosmid ends and plasmids (Table 1) also confirmed the substitutions and indels introduced in new CB4856 reference (Supporting Information).

From this combination of approaches we estimate that the resulting Hawaiian (HA) genome sequence totals 98,277,623 bases (98,291,416 including mitochondrial DNA), almost 2 Mb less than the N2 genome. All the CB4856 chromosomes are smaller than their N2 counterparts with II and V showing the greatest reductions at 2.57 and 3.54%, respectively (Table 2). The CB4856 genome size may be an underestimate because we have not been able to separate distinct copies of some repeats and some novel CB4856 sequences could still be missing. To provide a rough estimate of the total bases that might be missing from the CB4856 genome, we compared all the JR contigs to the CB4856 reference. There are only 22 assembled contigs (16,206 bases) that fail to align at all with our parameters. However, portions of other contigs fail to align to the CB4856 reference, and these total 1.12 Mb (2617 contigs). In addition, portions of other contigs (810 contigs, 1.74 Mb total) match the reference poorly (>1% of bases mismatched/inserted/deleted). Inspection of a sample of these contigs suggests that they could represent small insertions/deletions within the aligned sequence, duplicated sequence, poorly assembled sequence, or segments in the CB4856 with remaining problems. REAPR and ALE analysis also show some remaining problematic areas in the CB4856 reference, including high read coverage in regions that suggests unresolved duplications (sometimes overlapping more than one JR contig with poor alignment scores), low coverage, and inconsistent read pairing. Nonetheless, the CB4856 reference sequence

Table 4 SNV distribution

Chromosome	Total	Centers	SNV/kb centers	Arms	SNV/kb arms	% on arms
I	36,192	9,424	1.15	26,768	3.89	73.96
II	65,592	8,843	1.11	56,749	7.80	86.52
III	38,938	5,429	0.78	33,509	4.94	86.06
IV	36,198	9,016	1.00	27,182	3.20	75.09
V	129,096	18,179	2.02	110,917	9.30	85.92
X	21,034	8,076	1.08	12,958	1.27	61.61
Total	327,050	58,967	1.21	268,083	5.20	81.97

provides a much more complete picture of the genome for the analysis of variation.

Summary of sequence variation

To assess the sequence differences between the Hawaiian and N2 reference genomes, we aligned the CB4856 genome to N2 using LASTZ, a program that is able to perform pairwise, gapped, full chromosome-to-chromosome alignments even with quite distantly related sequences (Harris 2007). This alignment yielded a total of 327,050 SNVs and 79,529 indel events (File S1, File S2, File S3, File S4). In addition, 816 segments with sequence in both genomes failed to align with LASTZ and also failed to produce consistent alignments with different Smith–Waterman alignment implementations. Because these sequences fail to align, we cannot call variants in them. Inspection of several of these regions suggests several different origins. Some are extremely divergent and could represent overlapping deletions of an ancestral sequence, leaving unrelated sequences in the same place in the two present-day genomes. Others may just be of slightly greater divergence than LASTZ tolerates. Still others involve short sequences in one genome opposite a much larger region in the other genome that might be associated with repair events around a deletion. Because we were uncertain of the origin of the individual events, for purposes of analysis we treated the 816 segments as reciprocal insertion/deletion events.

If we treat the indels as deletions in N2 or CB4856 from a larger ancestral genome, we find similar numbers of deletion events in each genome, with small deletions (≤5 bases) accounting for 28,779 and 28,776 events in each N2 and CB4856, respectively. But the larger deletions (>5 bases) are slightly more numerous in CB4856 (12,181 *vs.* 9803), and the overall the number of bases deleted in CB4856 is much larger (2.2 Mb of ancestral genome lost in CB4856 *vs.* 0.92 Mb in N2). Chromosomes II and V have larger numbers of bases deleted. In addition to the indels, the 816 reciprocal insertion/deletion events show a similar trend of greater loss of sequence in the CB4856 genome, with these segments containing 458,526 bases in CB4856 and 1,109,331 bases in N2. *In toto* from the indels and reciprocal indels, N2 has 3.3 Mb of sequence not represented in CB4856, and CB4856 has 1.4 Mb of sequence not present in N2 (Table 3).

Among the deletions and insertions were copies of the transposons Tc1, Tc2, Tc3, Tc4, and Tc5. For example, among the 32 copies of Tc1 in the N2 reference, only 19 were detected in the CB4856 genome (Table S4). However, we found evidence for Tc1 copies in the CB4856 not present in N2. Because of the repeated nature of the sequence, the copies were incomplete in the CB4856 genome, but in 12 cases, Tc1 end sequences were present on both ends flanking a gap and, in another five cases, new Tc1 sequences were detected at one end. Differences in other Tc family transposons were also noted.

Our 327,050 SNV calls detect most of the SNVs reported earlier (Thompson *et al.* 2013; Vergara *et al.* 2014), but we find almost twice as many SNVs as Thompson *et al.* (2013) and 103,955 not reported by Vergara *et al.* (2014) (Figure 3). These novel SNVs are supported by conventional Sanger sequenced reads at a rate similar to those found in common by all three reports (Supporting Information). Many of the SNVs reported in Vergara *et al.* (2014) but not found here are adjacent to small deletions associated with homopolymer runs not reported in Vergara *et al.* (2014) [these authors report only 31,791 indels, compared with 46,544 detected in the Million Mutation Project (Thompson *et al.* 2013) and almost 80,000 here], suggesting that some of their called SNVs result from alignment issues and/or problems in runs (Becker *et al.* 2012).

Next, we looked at the distribution of the SNVs and indels across and within the chromosomes. *C. elegans* chromosomes have a distinctive organization, with the outer 20–30% of each chromosome (the arms) exhibiting a higher rate of recombination and a higher fraction of repeated sequences (Barnes *et al.* 1995; Rockman and Kruglyak 2009). They also contain the bulk of genes for large, rapidly evolving gene families. Consistent with previous reports, we find the number of SNVs is higher on the autosome arms than in the centers (Table 4, File S4, Figure S8), with as much as 86% of SNVs on arms . The distribution of indel events follows a similar pattern.

More strikingly, we noted a strong clustering of variants in smaller regions where we had extended sequence from the ends of aligned segments during the iterative alignment process (Figure 1A) and where we had replaced sequence with the JR-assembled contigs (Figure 1B). To detect these regions of higher divergence systematically, we clustered

Table 5 SNVs in divergent regions

	Divergent regions			Other regions		
Chromosome	SNVs	Bases	SNVs/kb	SNVs	Bases	SNVs/kb
I	3,940	87,170	45.20	32,252	14,985,253	2.15
II	27,649	709,991	38.94	37,943	14,569,354	2.60
III	13,962	344,847	40.49	24,976	13,438,853	1.86
IV	5,657	206,442	27.40	30,541	17,287,351	1.77
V	77,704	1,444,451	53.79	51,392	19,479,698	2.64
X	900	38261	23.52	20,134	17,680,605	1.14
Total	129,812	2,831,162	45.85	197,238	97,441,114	2.02

1-kb windows with >1.4% bases different or >500 bases deleted or inserted, retaining clusters of ≥9 kb. We manually reviewed each of the resulting clusters, adjusting the end points to more precisely reflect increasing SNV density, merging closely spaced clusters separated by repeats or other regions where SNVs were unable to be called. This procedure produced 61 regions containing a strikingly high proportion of SNVs and spanning 2,313,859 bases in the CB4856 sequence and 2,831,162 in the N2 genome (Table 5; see also Table S3 for a list of regions with coordinates and Table S5 for SNV counts in the absence of the divergent regions). The boundaries between these divergent regions and other regions are usually very sharp (Figure 2 and Figure 4). The segments are scattered across all six chromosomes, range in size from 8 to >162 kb and show 2–15.8% sequence divergence from N2 without an obvious correlation between size and divergence (Figure 5). The autosomal clusters fall principally on the arms with just one cluster in the central region of IV and three clusters in the central region of V (Figure S8). The X chromosome has only two clusters, and these have just 2 and 4.6% divergence each. The autosomal divergent regions include the *peel-1 zeel-1* region and the *glc-1* gene, both of which had been previously reported as having an elevated sequence divergence (Seidel *et al.* 2011; Ghosh *et al.* 2012).

Curiously, in comparing our results with prior studies using array CGH (Maydan *et al.* 2007; Maydan *et al.* 2010), we find that more than one-third of their deletion calls fall within these divergent regions, often extending across most of the region. Apparently the sequence divergence within the regions led to poor hybridization with the probes and resultant scoring of the area as deleted. The remaining arrayCGH deletions overlap extensively with deletions in the CB4856 reference except in one case [WBVar00091092; niDf71(III); chrIII:13778179–13781358] where we have normal coverage throughout the region.

Functional impact of the sequence variants

Prediction of the functional effects of the variants using WS230 annotation shows that a large portion of the protein-coding potential of the Hawaiian genome is altered (Table 6; File S5). Across the whole genome, 8140 (40%) of the 20,504 protein-coding genes have some coding change (for complete lists of the alterations, see File S2, File S3, and File S5, File S6). Of these genes, 1885 protein-coding genes have a likely loss-of-function (LOF) mutation—an induced stop codon, a frameshift, or a deletion across a splice junction—and 357 of these delete the gene entirely. Our reference also detects previously reported variants throughout the genome (Table S6).

In the 2.83 Mb of divergent regions, the relative impact of the variation on protein-coding genes is even greater. Of the 883 genes in these regions, 866 (98%) have some coding change, with 576 genes having LOF changes and 195 genes deleted entirely. Those genes with LOF changes could all be pseudogenes in CB4856, perhaps as a rapid means of adaptation (Olson 1999). However, the persistence of these regions over long evolutionary times (see below) suggests that they contain functionally important genes.

The genes altered by the LOF variants are disproportionately composed of members of large, rapidly evolving gene families, including the *math*, *bath* (btb-math), *clec* (c-lectin), *fbx* (F-box), and seven-transmembrane serpentine receptor genes. By contrast, the *nhr* (nuclear hormone receptor) genes are not overrepresented in LOF variants. The same large, rapidly evolving gene families are also heavily represented in the divergent regions. The *math* family is notable, having 37 of its 49 members in the divergent regions. Within the divergent regions the *math*, *bath*, and *fbx* families also suffer a disproportionate number of LOF variants. The members of these families in the divergent regions that do not contain LOF variants also show a d_N/d_S ratio (ratio of divergence at nonsynonymous and synonymous sites) approaching or even exceeding 1. By contrast, the seven-transmembrane and *nhr* gene families are relatively spared. Members of these families without LOF variants also show a lower fraction of nonsynonymous changes than the *math* and *fbx* families.

Some genes with easily visible mutant phenotypes contain LOF variants, such as *ced-1*, *ced-6*, *unc-13*, and *unc-49*. However, inspection of the variants in these genes in comparison with WormBase models and the extensive modENCODE RNA-seq data (Gerstein *et al.* 2014) suggest in each case that the annotation likely requires correction. For example, the putative stop codon in the *unc-13* model falls in a splice form that has no RNA-seq support; instead, the RNA-seq evidence suggests that the exon in question is an alternate first exon and the "nonsense" change lies in the 5′ UTR. Similarly, the putative frameshift mutation (an insertion of

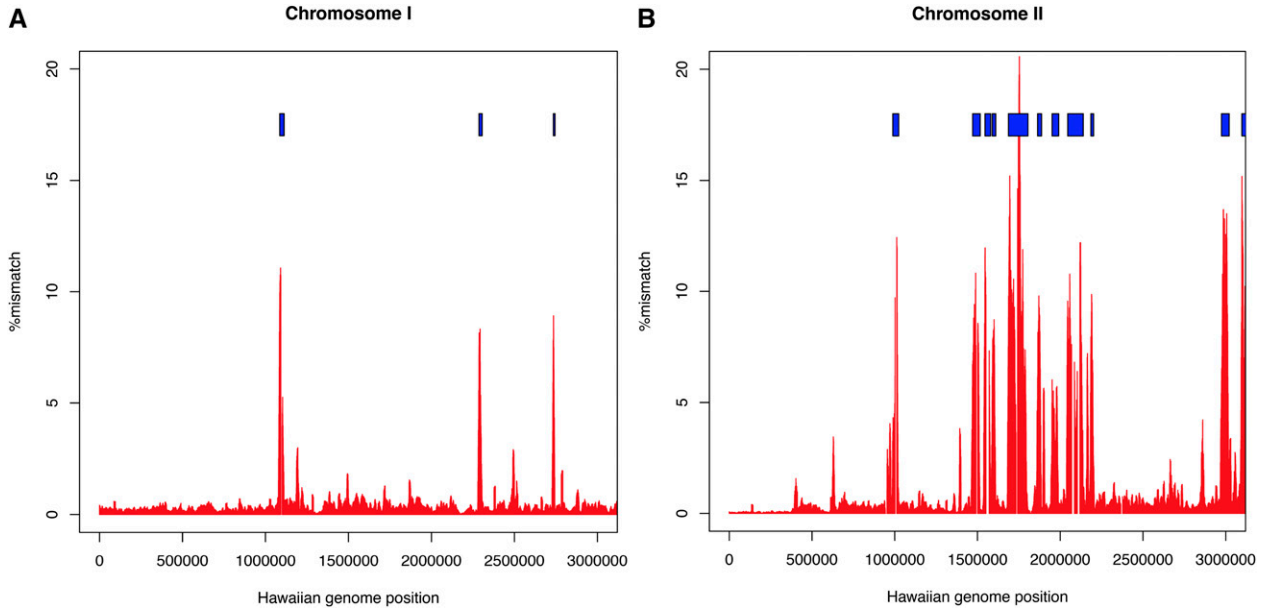


Figure 4 Density of variant sites in the first three megabases of (A) chromosome I and (B) chromosome II. Blue boxes indicate the regions identified as highly divergent.

a G) in *ced-6* falls adjacent to a noncanonical splice donor sequence in the WormBase gene model, where the presence of an extra G would allow the use of a canonical GT donor. Inspection of data from an N2 strain (VC2010) suggests that a G is missing here in the N2 reference.

Other regions that include genes such as *tbx-30*, *tbx-31*, *vit-3*, and *vit-4* have suffered large deletions in the CB4856 genome, with *tbx-30* deleted entirely. However, in the N2 genome *tbx-30* has an exact duplicate, *tbx-42*, as part of an inverted duplicated segment; presumably, the loss of a single copy is tolerated. The two *vit* genes are adjacent in the genome, and in this case the deletion appears to fuse the two genes into a single gene.

Although genes have been lost in the CB4856 genome, genes may also be present in CB4856 that are defective or absent in N2. To look for these genes, we generated gene models by using an *ab initio* gene prediction tool, Genefinder, and by aligning CB4856 RNA-seq data (Table 1) against the CB4856 reference (Trapnell *et al.* 2012). Among the recent N2 pseudogenes that arose by duplication, we found two gene models in CB4856 that had full opening reading frames that were similar to the parent genes. In sequences inserted in CB4856 relative to N2, Genefinder produced models that were often supported by RNA-seq data and had similarity to members of multigene families. For example, two large insertions of 5 and 15 kb in one divergent region (II:1544781–1579478 CB4856 coordinates; II:1613411–1636156 N2 coordinates) contain seven Genefinder models, all of which have similarity to *fbx* genes. Similarly, in regions where LASTZ failed to align the N2 and CB4856 sequence, there were gene models with RNA-seq support in the CB4856 sequences. For example, in two adjacent regions that fail to align by LASTZ in a divergent region (II:2974336–3020963 CB4856 coordinates; II:3197502–3250769 N2 coordinates), there are two gene models, one with similarity to seven-transmembrane receptors and the second to *fbx* genes. Thus, while CB4856 has lost genes present in N2, it has also gained genes.

We also examined genes present in both species with multiple differences that suggested that they were inactive in CB4856. One gene—*F47H4.2*—stood out because of the multiple changes reported from the LASTZ alignments, including 691 SNVs across the two isoforms that, considered together, result in 511-amino-acid substitutions, 1 nonsense mutation, and 13 frameshifting indels. Despite these multiple variants, a Genefinder model in the region of CB4856 predicts a protein of 628 amino acids with its seven exons having open reading frames similar in length to those in N2 and with six of those having splice junctions in precisely equivalent places (Figure S9). The predicted nonsense codon is present in an unused frame, a consequence of flanking, compensating frameshift variants. The 628-amino-acid predicted protein, when evaluated by blastp (Altschul *et al.* 1990) against the N2 proteins, yielded a match that covered amino acids 1–633 (P = 2.7*e*-127) of an *F47H4.2* isoform, spanning two FTH domains that are also found in F-box genes. The Genefinder exons also are matched by RNA-seq reads. Open reading frames also exist in CB4856 for the final three exons of *F47H4.2* and are partially incorporated into a second Genefinder model (Figure S9). These results suggest that, rather than containing an inactivated gene, the CB4856 region encodes a homologous protein. Like *F47H4.2*, other genes that appear to have suffered a LOF variant in CB4856 also have Genefinder models in syntenic regions

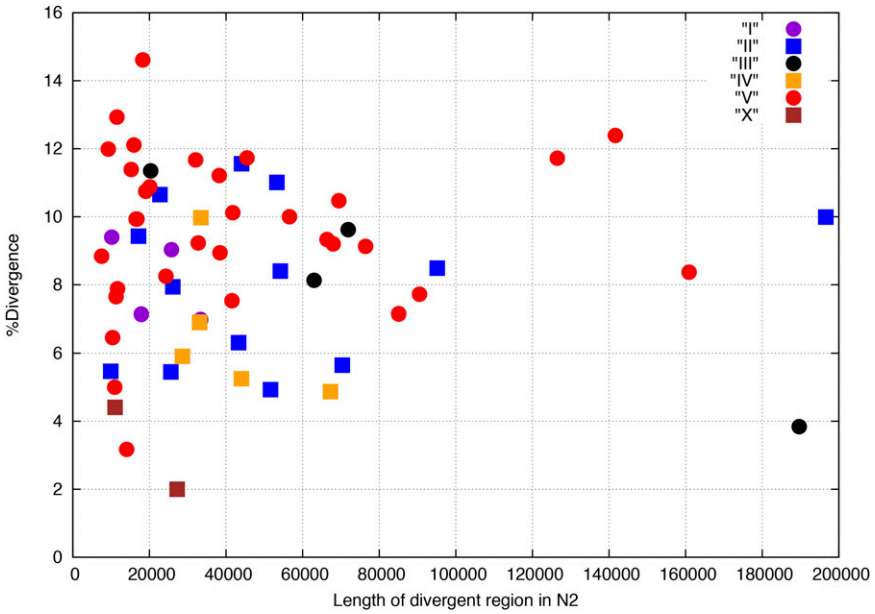


Figure 5 Percent divergence by length of divergent region per chromosome. The mutational events (SNVs and indels, counting each indel as a single event) per aligned bases (percentage divergence) are plotted for each region against the length of the region in N2. The chromosomal assignment for each region is indicated in the inset.

that are often supported by RNA-seq data and may similarly encode full-length proteins.

Divergent regions in other C. elegans strains

Prior studies of the *glc-1* and *peel-1 zeel-1* regions showed that both CB4856 and N2 haplotypes were widely distributed in strains recovered from the wild (Seidel *et al.* 2011; Ghosh *et al.* 2012). To determine haplotype representation of the 61 divergent regions in other wild strains, we exploited the MPS sequence reads acquired previously (Thompson *et al.* 2013) for each of 39 strains and aligned them against the N2 reference and the CB4856 reference. We expected that, if a particular region in a given strain had the N2 haplotype, its reads would align well with the N2 reference but poorly with the CB4856 reference and vice versa. To assess the alignment quality, we calculated ALE scores using the N2 and CB4856 reads aligned against each other as controls.

Using the ALE scores (Figure S10), we cataloged the regions as N2-like, CB4856-like, intermediate, or different from either. The results (Figure 6) for the 44 regions giving consistent scoring show that, while, overall, N2-like haplotypes are most frequent, the CB4856-like sequence is found in at least one other strain for all but six of the regions. For five regions, the CB4856-like haplotype is predominant. Two regions are represented mainly by sequence that matches neither N2 nor CB4856 well, suggesting the presence of another version of the sequence in these strains. Other regions match both N2 and CB4856 at intermediate levels, *e.g.*, V:18193641–18260001, and inspection of these regions suggests that some have one segment that matches CB4856 well and a second segment that matches N2 well, perhaps reflecting recombination events between the two haplotypes. Most of the strains have unique combinations of sequences; however, JU1171, MY2, and MY14 all share the same pattern (both MY2 and MY14 were isolated from Munster, Germany; share 96% of SNV calls and share 96% with JU1171, isolated in Chile; they likely represent a single isotype), and ED3057 and ED3072 are similar (the latter two were both isolated in Kenya and share 97% of their SNV calls and also likely represent a single isotype).

Discussion

We have used a variety of resources and methods to produce a draft reference CB4856 genome sequence. We combined iterative alignment of deep whole-genome sequence reads to a progressively corrected CB4856 reference version with a *de novo* whole-genome assembly. We used the assembly assessment tools ALE and REAPR to monitor progress, to identify problem areas needing improvement, and to break suspicious joins in the *de novo* assembly. Sequence reads from ILs and RILs helped guide placement of the *de novo* contigs into the reference. Sanger reads from fosmid-insert ends and random clones were critical in validating the reference sequence at each step.

The resulting CB4856 reference sequence extends and refines the scope of the variation between N2 and CB4856 from prior studies. In particular, the draft sequence reveals 61 regions of substantially higher divergence than the rest of the genome. These regions total 2.8% of the N2 genome and contain 40% (129,812/327,050) of total SNVs and 21% (16,822/79,558) of total indels. A survey of genome sequence data from other wild isolates of *C. elegans* shows that generally in these regions they closely resemble one or the other genome. However, some isolates appear to have regions that are divergent from both N2 and CB4856 and in segments outside the divergent regions some strains show divergence from both N2 and CB4856. These findings

Table 6 Genes in diverged regions and with LOF mutations

	Genome			Divergent regions			
Gene class	Total	Disabling	Expected[a]	Total	Expected	Disabling	Expected[b]
Serpentine receptor (*sr*/str*)	1346	204	123.7 2.20e-13[c]	118	49.7 7.38e-14	72	80.0 8.70e-1
F-box (*fbx**)	353	129	32.5 1.56e-45	71	15.3 3.75e-28	60	46.3 1.51e-4
C-lectin (*clec*)	254	53	23.4 1.05e-8	31	11.0 1.81e-7	19	20.2 7.49e-1
Math (*math*)	48	39	4.4 1.92e-32	37	2.1 2.00e-41	34	24.1 1.47e-4
Bath (*bath*)	37	16	3.4 4.83e-8	15	1.6 1.11e-11	14	9.7 1.42e-2
Nuclear hormone receptor (*nhr*)	278	26	25.6 4.90e-1	27	11.8 7.38e-05	10	17.6 9.99e-1

[a] The expected number of disabled genes in the total genome based on 20,504 genes and 1885 disabled overall.
[b] The expected number of genes in the divergent regions based on 883 genes of the 20,504 genes in the genome and 576 of the 883 genes disabled.
[c] Hypergeometric test.

suggest that N2 and CB4856 do not capture all the divergent haplotypes in these regions or even all the regions with divergent haplotypes extant in this species. The reference genome, associated data sets, and annotations can be viewed through a track data hub listed on the University of California at Santa Cruz browser (http://genome.ucsc.edu/cgi-bin/hgHubConnect and connect to "*C. elegans* isolates").

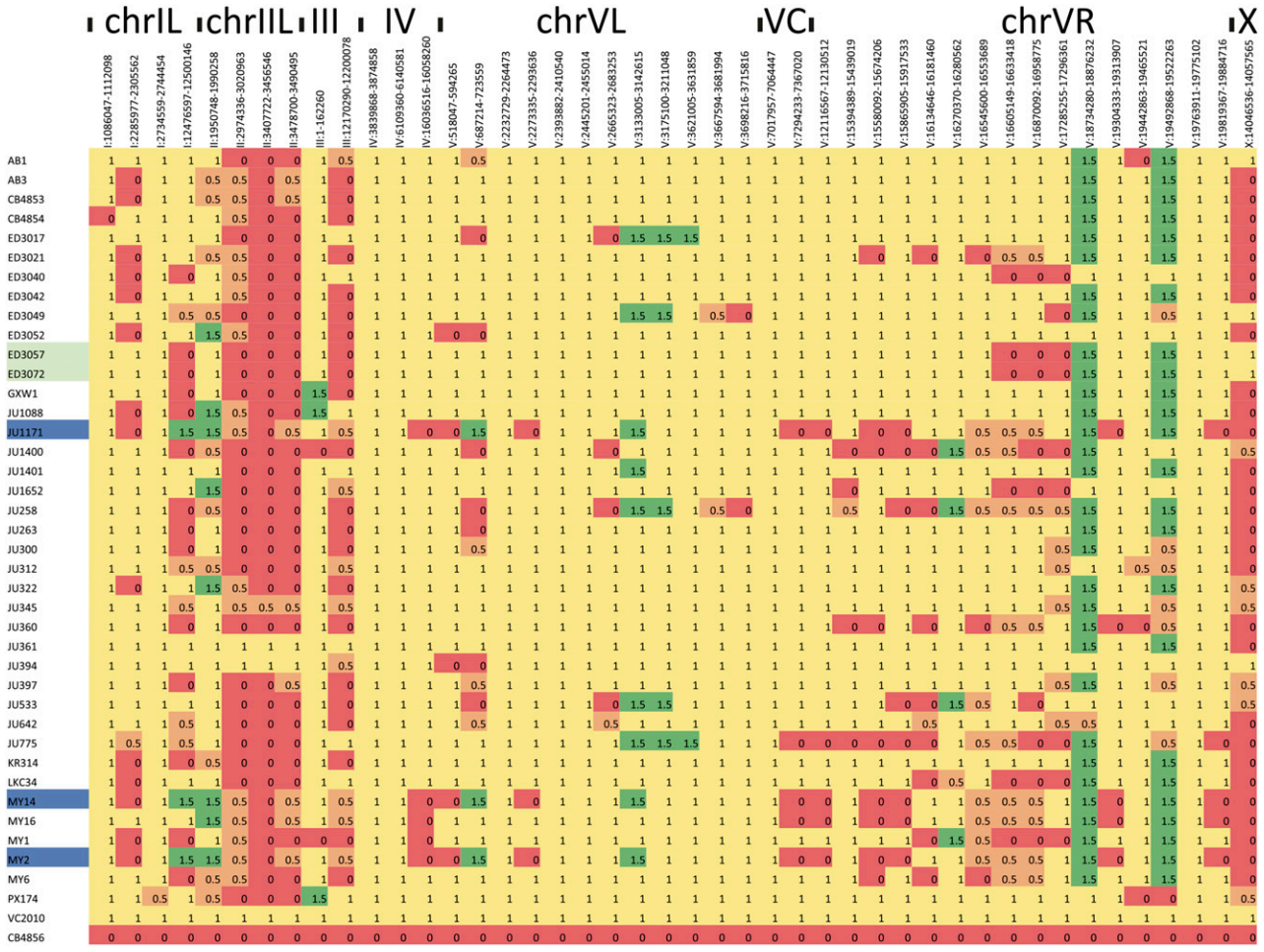


Figure 6 A heatmap representation of the allelic content of the 39 strains (rows) across 44 of the 61 divergent regions (columns). Regions matching N2 (yellow) and CB4856 (red) are indicated along with intermediate regions (orange) and regions different from either (green). For reference, an N2-derived strain, VC2010, and CB4856 are shown in the bottom two rows. Strains that may represent the same isotype are highlighted in blue and green.

The high level of divergence between the two haplotypes of N2 and CB4856 in these regions indicates an ancient origin. Possible explanations could include long-term balancing selection, as suggested for the *glc-1* haplotypes (Ghosh *et al.* 2012), that were estimated to have arisen $\sim 7 \times 10^6$ generations ago. Applying the same methods to the 61 regions described here gives an overall estimate of d_S (average number of nucleotide differences between sequences per synonymous site) of 0.12 and similar estimate of 6.7×10^6 generations. The male–female species *C. remanei* shows 4% divergence in wild populations (Dey *et al.* 2012). Perhaps these divergent regions are a remnant of the variation present at the conversion to a hermaphroditic species, with all 61 regions maintained by balancing selection since the origin of the species. Alternatively, the sequences may have evolved independently in different, isolated populations, followed by one or more subsequent crossbreeding events, leading to admixture and introgression. Regardless of the origin, haplotypes for only a few regions have persisted, perhaps maintained by balancing selection. Although the high degree of sequence divergence might be expected to interfere with crossing over in the initial hybrid, resulting in a severe lack of fitness, nematodes with this level of divergence have been found to be fertile (Dey *et al.* 2013). Perhaps the need for only one recombination event per chromosome makes the *Caenorhabditis* species more tolerant of long tracts of divergent sequences. Regardless of the nature of the original event(s), balancing selection is likely to be playing a role in the maintenance of these regions across the globe.

What is driving the balancing selection? Genes that belong to large, rapidly evolving gene families and that have a putative role in the interaction with the environment are abundant in these regions (Thomas 2006; Thomas and Robertson 2008). The alternative haplotypes may harbor combinations of genes and alleles that provide selective advantage in the face of changing environmental conditions. The *glc-1* polymorphisms provide a specific example of polymorphisms that would confer selective advantage in an environment containing avermectin or related compounds. In the absence of avermectin, the defective *glc-1* may lead to a fitness disadvantage (Ghosh *et al.* 2012). But also, *glc-1* lies in a divergent region with multiple seven-transmembrane receptor and *clec* genes. Perhaps these other genes also have a role in balancing selection.

Beyond the revelation of these divergent regions, the CB4856 reference sequence provides investigators with a comprehensive list of the changes between CB4856 and N2 (File S1, File S2, File S3, File S5). These lists should prove useful to those investigators using wild isolates of *C. elegans*. For example, the failure of a probe to hybridize well to the CB4856-derived sequences might lead to false negatives. RNA-seq biases introduced by mapping to the N2 reference may also distort any signal. These biases are particularly relevant for the divergent regions. The extensive catalog of differences now available between these two strains in combination with the powerful genetic approaches available in *C. elegans* should facilitate the dissection of the growing number of phenotypic differences.

Our findings also have broader implications for studies comparing genome sequences to reference sequences using short reads. The high degree of sequence divergence in these divergent regions compromises alignment of short reads. Studies that use only standard alignment of reads to a reference fail to assess such divergent sequence. If similar areas of high divergence are present in other species, they too would be missed. Our results indicate the importance of going beyond simple read alignment in the assessment of variability between different genomes.

Acknowledgments

We thank Brent Ewing and Stephane Flibotte for many discussions on sequence alignment, assembly, and analysis; Yeh Teng Wen and the JR-Assembler team for help in installing and running their assembler; and Asher Cutter for ideas about the origins of the divergent regions. The work was supported by *The American Recovery and Reinvestment Act* Grand Opportunities (ARRA GO) grant HG005921 from the *National Human Genome Research Institute (NHGRI)*, by grant HG007355 from NHGRI, and by the William H. Gates Chair of Biomedical Sciences. L.B.S. was funded by The Netherlands Organisation for Scientific Research (project no. 823.01.001).

Literature Cited

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Andersen, E. C., J. P. Gerke, J. A. Shapiro, J. R. Crissman, R. Ghosh *et al.*, 2012 Chromosome-scale selective sweeps shape *Caenorhabditis elegans* genomic diversity. Nat. Genet. 44: 285–290.

Andersen, E. C., J. S. Bloom, J. P. Gerke, and L. Kruglyak, 2014 A variant in the neuropeptide receptor npr-1 is a major determinant of *Caenorhabditis elegans* growth and physiology. PLoS Genet. 10: e1004156.

Andersen, E. C., T. C. Shimko, J. R. Crissman, R. Ghosh, J. S. Bloom *et al.*, 2015 A powerful new quantitative genetics platform, combining *Caenorhabditis elegans* high-throughput fitness assays with a large collection of recombinant strains. G3 (Bethesda) 5: 911–920.

Barnes, T. M., Y. Kohara, A. Coulson, and S. Hekimi, 1995 Meiotic recombination, noncoding DNA and genomic organization in *Caenorhabditis elegans*. Genetics 141: 159–179.

Becker, E. A., C. M. Burns, E. J. Leon, S. Rajabojan, R. Friedman *et al.*, 2012 Experimental analysis of sources of error in evolutionary studies based on Roche/454 pyrosequencing of viral genomes. Genome Biol. Evol. 4: 457–465.

Cao, J., K. Schneeberger, S. Ossowski, T. Gunther, S. Bender *et al.*, 2011 Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat. Genet. 43: 956–963.

Capra, E. J., S. M. Skrovanek, and L. Kruglyak, 2008 Comparative developmental expression profiling of two *C. elegans* isolates. PLoS ONE 3: e4055.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science 282: 2012–2018.

Chu, T. C., C. H. Lu, T. Liu, G. C. Lee, W. H. Li *et al.*, 2013 Assembler for de novo assembly of large genomes. Proc. Natl. Acad. Sci. USA 110: E3417–E3424.

Clark, S. C., R. Egan, P. I. Frazier, and Z. Wang, 2013 ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 29: 435–443.

de Bono, M., and C. I. Bargmann, 1998 Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. Cell 94: 679–689.

Dey, A., Y. Jeon, G. X. Wang, and A. D. Cutter, 2012 Global population genetic structure of *Caenorhabditis remanei* reveals incipient speciation. Genetics 191: 1257–1269.

Dey, A., C. K. Chan, C. G. Thomas, and A. D. Cutter, 2013 Molecular hyperdiversity defines populations of the nematode *Caenorhabditis brenneri*. Proc. Natl. Acad. Sci. USA 110: 11056–11060.

Doroszuk, A., L. B. Snoek, E. Fradin, J. Riksen, and J. Kammenga, 2009 A genome-wide library of CB4856/N2 introgression lines of *Caenorhabditis elegans*. Nucleic Acids Res. 37: e110.

Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe *et al.*, 2011 Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477: 419–423.

Ghosh, R., E. C. Andersen, J. A. Shapiro, J. P. Gerke, and L. Kruglyak, 2012 Natural variation in a chloride channel subunit confers avermectin resistance in *C. elegans*. Science 335: 574–578.

Gerstein, M. B., J. Rozowsky, K. K. Yan, D. Wang, C. Cheng *et al.*, 2014 Comparative analysis of the transcriptome across distant species. Nature 28: 445–448.

Harris, R. S., 2007 *Improved Pairwise Alignment of Genomic DNA*. Ph.D. Thesis, The Pennsylvania State University Press.

Hodgkin, J., and T. Doniach, 1997 Natural variation and copulatory plug formation in *Caenorhabditis elegans*. Genetics 146: 149–164.

Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman *et al.*, 2013 REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14: R47.

Kammenga, J. E., A. Doroszuk, J. A. Riksen, E. Hazendonk, L. Spiridon *et al.*, 2007 A *Caenorhabditis elegans* wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3. PLoS Genet. 3: e34.

Kent, W. J., R. Baertsch, A. Hinrichs, W. Miller, and D. Haussler, 2003 Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. USA 100: 11484–11489.

Koch, R., H. G. van Luenen, M. van der Horst, K. L. Thijssen, and R. H. Plasterk, 2000 Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*. Genome Res. 10: 1690–1696.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. PLoS Genet. 2: e222.

Mackay, T. F., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles *et al.*, 2012 The *Drosophila melanogaster* Genetic Reference Panel. Nature 482: 173–178.

Maydan, J. S., S. Flibotte, M. L. Edgley, J. Lau, R. R. Selzer *et al.*, 2007 Efficient high-resolution deletion discovery in *Caenorhabditis elegans* by array comparative genomic hybridization. Genome Res. 17: 337–347.

Maydan, J. S., A. Lorch, M. L. Edgley, S. Flibotte, and D. G. Moerman, 2010 Copy number variation in the genomes of twelve natural isolates of *Caenorhabditis elegans*. BMC Genomics 11: 62.

Nicholas, W. L., E. C. Dougherty, and E. L. Hansen, 1959 Axenic cultivation of *Caenorhabditis briggsae* (Nematoda, Rhabditidae) with chemically undefined supplements: comparative studies with related nematodes. Ann. N. Y. Acad. Sci. 77: 218–236.

Olson, M. V., 1999 When less is more: gene loss as an engine of evolutionary change. Am. J. Hum. Genet. 64: 18–23.

Perkins, J. D., 2010 *Comparison of Fosmid Libraries Made from Two Geographic Isolates of Caenorhabditis elegans*. M.Sc. Thesis, University of British Columbia, Vancouver.

Pollard, D. A., and M. V. Rockman, 2013 Resistance to germline RNA interference in a *Caenorhabditis elegans* wild isolate exhibits complexity and nonadditivity. G3 (Bethesda) 3: 941–947.

Rockman, M. V., and L. Kruglyak, 2008 Breeding designs for recombinant inbred advanced intercross lines. Genetics 179: 1069–1078.

Rockman, M. V., and L. Kruglyak, 2009 Recombinational landscape and population genomics of *Caenorhabditis elegans*. PLoS Genet. 5: e1000419.

Rockman, M. V., S. S. Skrovanek, and L. Kruglyak, 2010 Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. Science 330: 372–376.

Schacherer, J., J. A. Shapiro, D. M. Ruderfer, and L. Kruglyak, 2009 Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. Nature 458: 342–345.

Schneeberger, K., S. Ossowski, F. Ott, J. D. Klein, X. Wang *et al.*, 2011 Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc. Natl. Acad. Sci. USA 108: 10249–10254.

Schwartz, S., W. J. Kent, A. Smit, Z. Zhang, R. Baertsch *et al.*, 2003 Human-mouse alignments with BLASTZ. Genome Res. 13: 103–107.

Seidel, H. S., M. Ailion, J. Li, A. van Oudenaarden, M. V. Rockman *et al.*, 2011 A novel sperm-delivered toxin causes late-stage embryo lethality and transmission ratio distortion in *C. elegans*. PLoS Biol. 9: e1001115.

Snoek, L. B., K. J. Van der Velde, D. Arends, Y. Li, A. Beyer *et al.*, 2013 WormQTL: public archive and analysis web portal for natural variation data in *Caenorhabditis* spp. Nucleic Acids Res. 41: D738–D743.

Snoek, L. B., H. E. Orbidans, J. J. Stastna, A. Aartse, M. Rodriguez *et al.*, 2014 Widespread genomic incompatibilities in *Caenorhabditis elegans*. G3 (Bethesda) 4: 1813–1823.

Sterken, M. G., L. B. Snoek, J. E. Kammenga, and E. C. Andersen, 2015 The laboratory domestication of *Caenorhabditis elegans*. Trends Genet. 31: 224–231.

Stewart, M. K., N. L. Clark, G. Merrihew, E. M. Galloway, and J. H. Thomas, 2005 High genetic diversity in the chemoreceptor superfamily of *Caenorhabditis elegans*. Genetics 169: 1985–1996.

Stoeckius, M., D. Grun, and N. Rajewsky, 2014 Paternal RNA contributions in the *Caenorhabditis elegans* zygote. EMBO J. 33: 1740–1750.

Sulston, J. E., and S. Brenner, 1974 The DNA of *Caenorhabditis elegans*. Genetics 77: 95–104.

Swan, K. A., D. E. Curtis, K. B. McKusick, A. V. Voinov, F. A. Mapa *et al.*, 2002 High-throughput gene mapping in *Caenorhabditis elegans*. Genome Res. 12: 1100–1105.

Thomas, J. H., 2006 Adaptive evolution in two large families of ubiquitin-ligase adapters in nematodes and plants. Genome Res. 16: 1017–1030.

Thomas, J. H., and H. M. Robertson, 2008 The *Caenorhabditis* chemoreceptor gene families. BMC Biol. 6: 42.

Thompson, O., M. Edgley, P. Strasbourger, S. Flibotte, B. Ewing *et al.*, 2013 The million mutation project: a new approach to genetics in *Caenorhabditis elegans*. Genome Res. 23: 1749–1762.

Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim *et al.*, 2012 Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7: 562–578.

van der Velde, K. J., M. de Haan, K. Zych, D. Arends, L. B. Snoek *et al.*, 2014 WormQTLHD: a web database for linking human disease to natural variation data in *C. elegans*. Nucleic Acids Res. 42: D794–D801.

Vergara, I. A., M. Tarailo-Graovac, C. Frech, J. Wang, Z. Qin *et al.*, 2014 Genome-wide variations in a natural isolate of the nematode *Caenorhabditis elegans*. BMC Genomics 15: 255.

Vinuela, A., L. B. Snoek, J. A. Riksen, and J. E. Kammenga, 2012 Aging uncouples heritability and expression-QTL in *Caenorhabditis elegans*. G3 (Bethesda) 2: 597–605.

Volkers, R. J., L. B. Snoek, C. J. Hubar, R. Coopman, W. Chen *et al.*, 2013 Gene-environment and protein-degradation signatures characterize genomic and phenotypic diversity in wild *Caenorhabditis elegans* populations. BMC Biol. 11: 93.

Wicks, S. R., R. T. Yeh, W. R. Gish, R. H. Waterston, and R. H. Plasterk, 2001 Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. Nat. Genet. 28: 160–164.

Communicating editor: M. Johnston

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1

Remarkably Divergent Regions Punctuate the Genome Assembly of the *Caenorhabditis elegans* Hawaiian Strain CB4856

Owen A. Thompson, L. Basten Snoek, Harm Nijveen, Mark G. Sterken, Rita J. M. Volkers, Rachel Brenchley, Arjen van't Hof, Roel P. J. Bevers, Andrew R. Cossins, Itai Yanai, Alex Hajnal, Tobias Schmid, Jaryn D. Perkins, David Spencer, Leonid Kruglyak, Erik C. Andersen, Donald G. Moerman, LaDeana W. Hillier, Jan E. Kammenga, and Robert H. Waterston

DOI: 10.1534/genetics.115.175950

Supplementary Figures

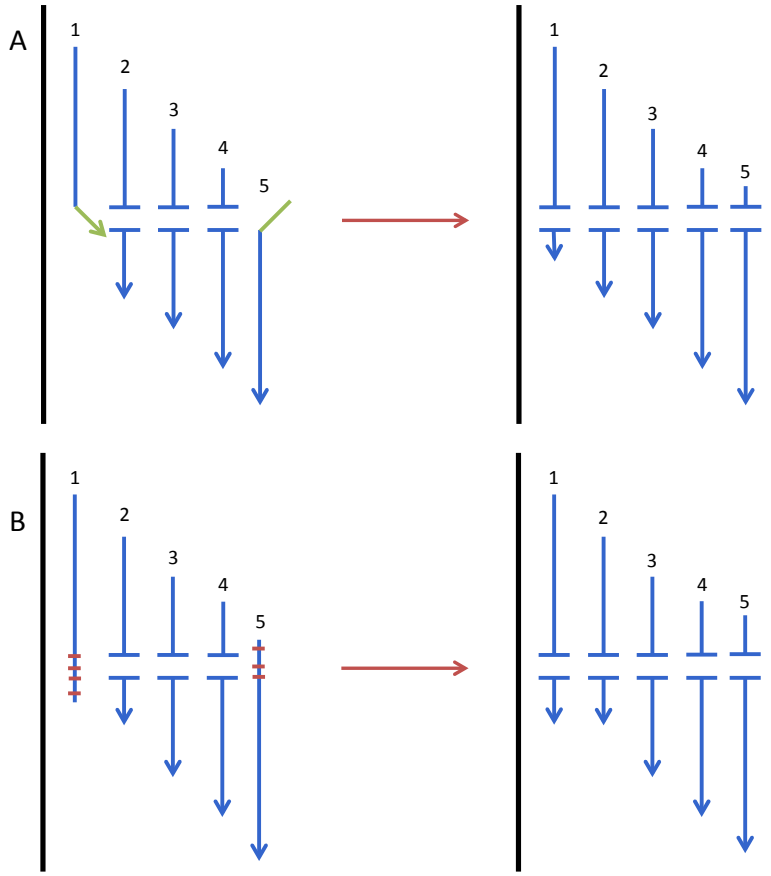


Figure S1 Read realignment via CIGAR modification using information from adjacent reads. (a) Elimination of Soft Clipping: A deletion with respect to the reference sequence is called in a fraction of reads at a given site, but induces some right-sided clipping (read 1, clipped sequence in green) or left-sided clipping (read 5) when reads cannot sufficiently frame the variant. Clipped reads may also occur when encountering high SNV density or compound variants. Testing the sequence of the read against the variant sequence predicted by other alignments to the same locus allows for extension of clipped alignments through the variant region, reducing the number of spurious clipping sites. (b) Variant regions near the ends of reads may also cause aligners to generate erroneous unclipped alignments, often preferring the introduction of multiple SNVs to the introduction of a gap (red tick marks, reads 1 and 5), as has been noted many times elsewhere. In both figure (a) and figure (b), analyzing each read for agreement with variant alignments at the same locus allows us to increase support for the predominant variant call(s) at the locus in question when consensus thresholds are implemented.

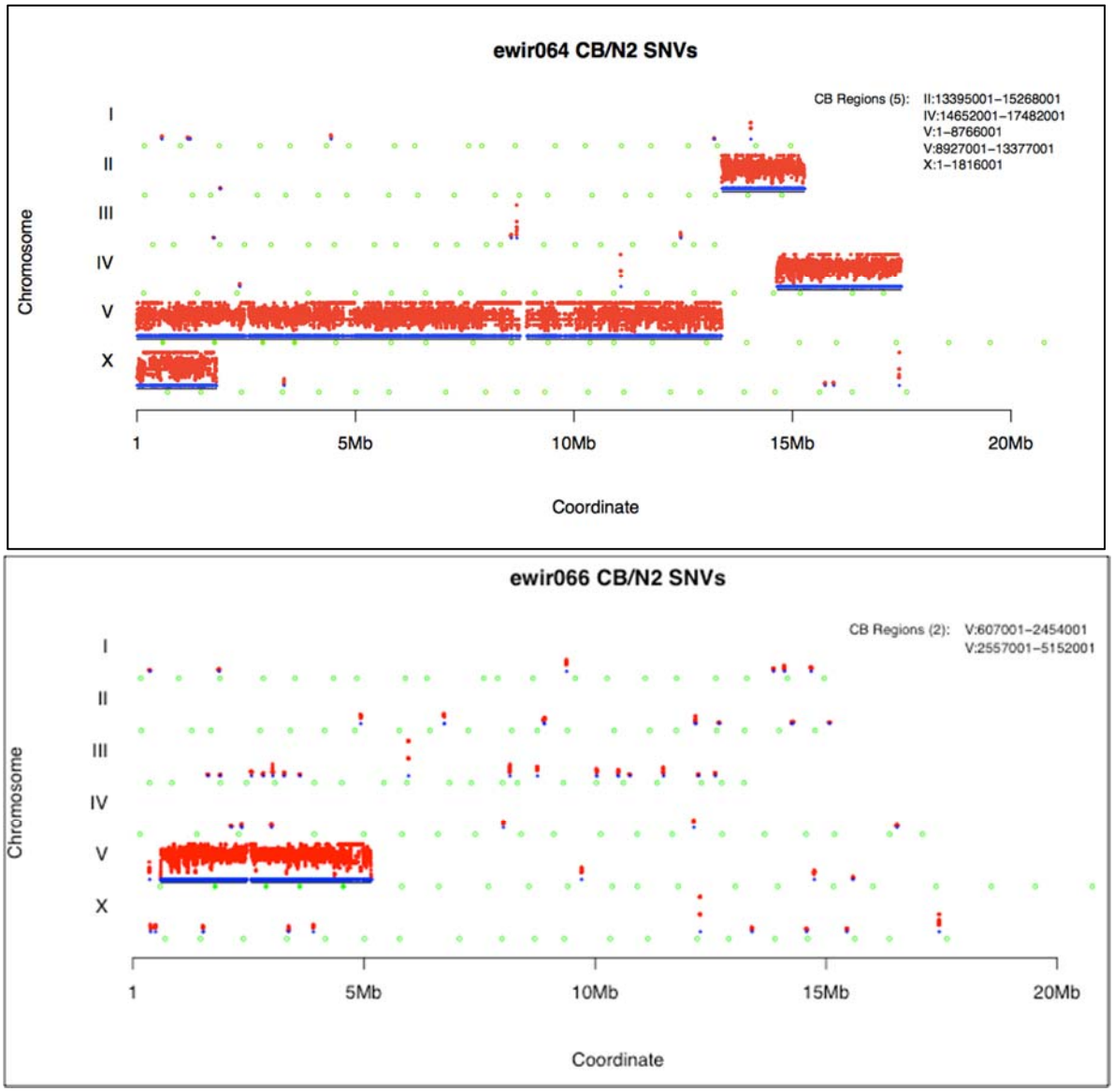


Figure S2 Example RIL CB/N2 sequence contribution. We detected CB4856 contributions to the genome of the introgressed N2/CB4856 line ewir064 and ewir066 by scanning each chromosome in bins of 10kb to determine the fraction of known CB4856 SNPs. The only shared regions in the two strains are on V:607001-5152001, thus locating any contig with matches to these strains to this interval. (Blue lines = CB4856 SNPs detected, vertical displacement of red coordinates for each chromosome = fraction of known CB4856 SNPs detected in the window, Empty green circles = PCR negative for a known variant, Filled green circles = PCR positive for a known variant (Doroszuk *et al.* 2009)).

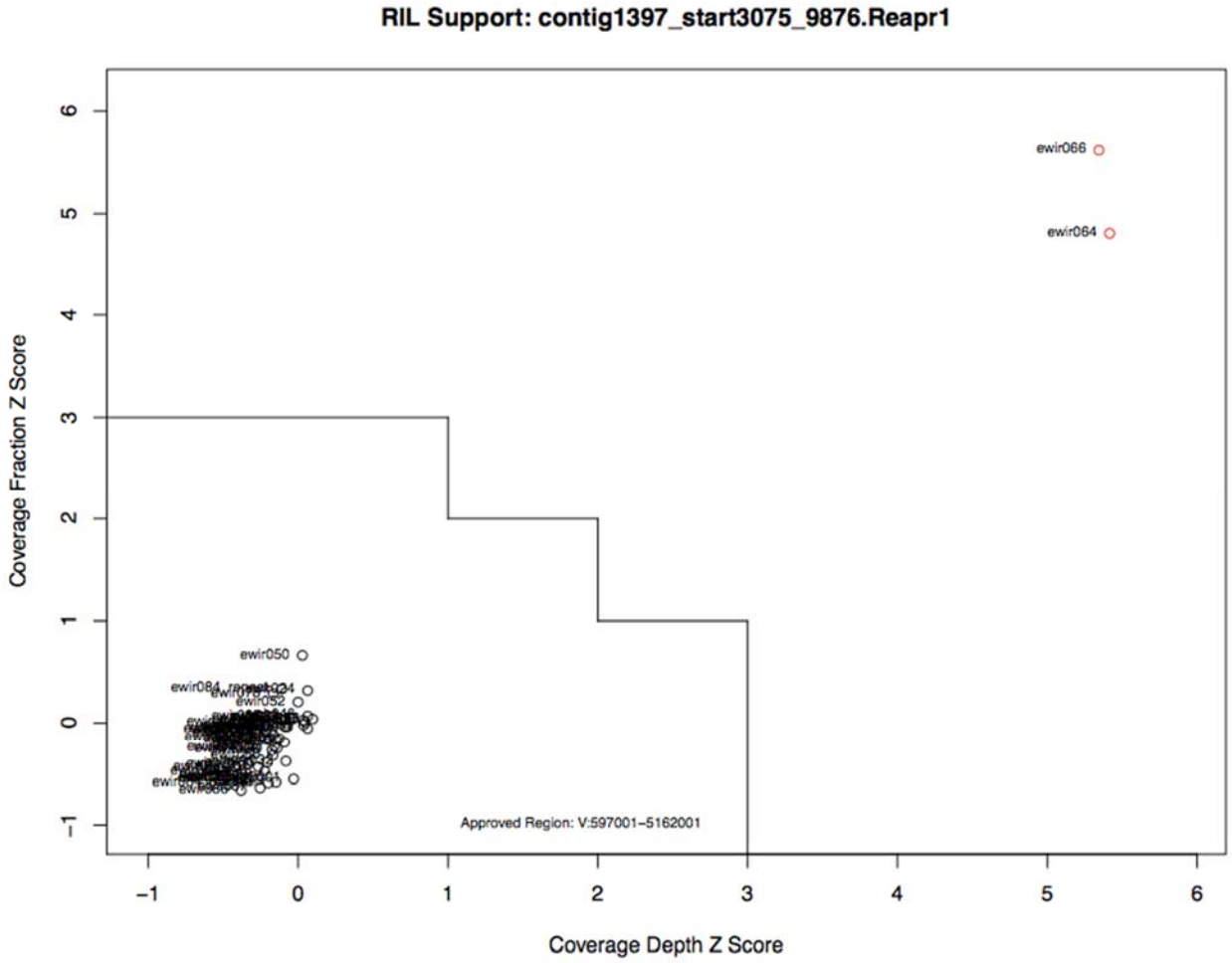


Figure S3 IL Usage to Map JR-Contig 1397 to chromosome V. SOLiD reads from the 49 IL lines were aligned to contig1397_start3075_9876.Reapr1 and Z scores calculated in each strain for both the depth of coverage (X axis) and the fraction of bases producing that coverage (Y axis), roughly approximating the levels of CB4856-unique sequence and the overall agreement with CB4856, respectively. IL lines ewir064 and ewir066 are substantially elevated in each statistic, indicating that these two ILs have sequence from this assembly contig. The shared CB4856 regions defined previously (e.g. in Figure S3) locate the contig to a genomic region.
(waterston.gs.washington.edu/trackhubs/isolates/JR-Assembler/images)

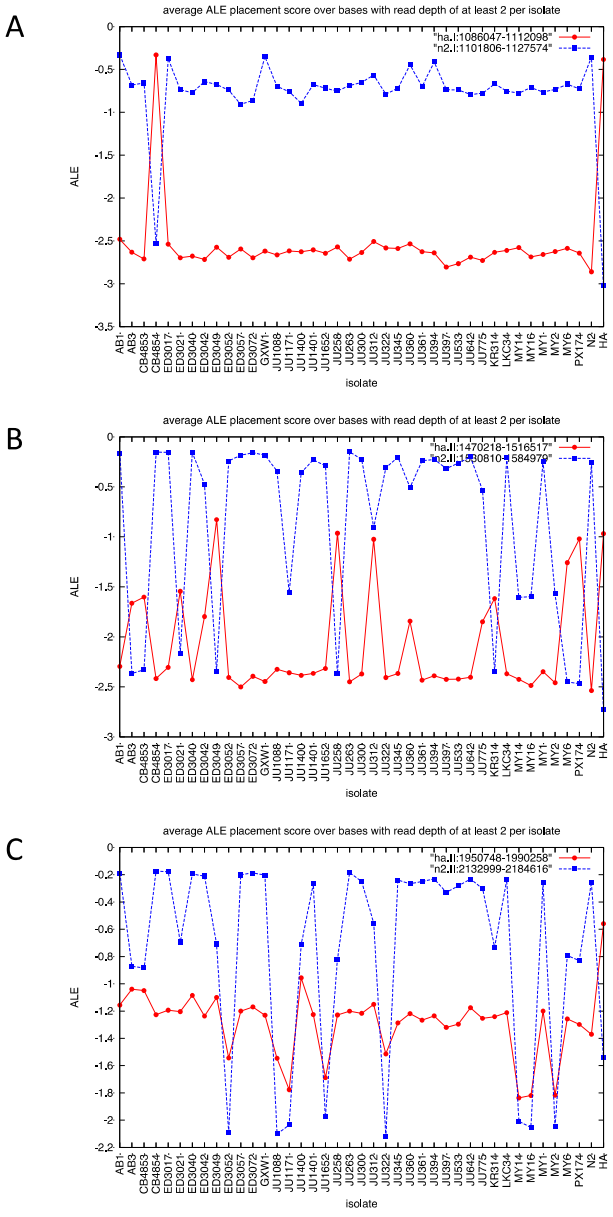


Figure S4 Match of sequence reads from 39 different wild isolates, N2 and CB4856 (HA) against the N2 genome and the CB4856 genome in a divergent region from chromosome I. The extent of the match is measured by ALE placement scores averaged across the region, with larger negative scores indicating a poorer fit. In (A) this region only the fourth strain (CB4854) resembles the CB4856 sequence. In (B) most strains are similar to the N2 pattern, but ED3049(9), JU258(19), MY6(38) and PX174(39) resemble HA. Others show an intermediate match, e.g. AB3(2) and CB4853(3). In (C), most strains again resemble N2, none resemble HA, a few are intermediate, e.g., AB3(2) and CB4853(3), and others are different from both N2 and HA, e.g., ED3052(10), JU1088(14) and JU1171(15). Figure S11 contains these same data for all of the divergent regions.

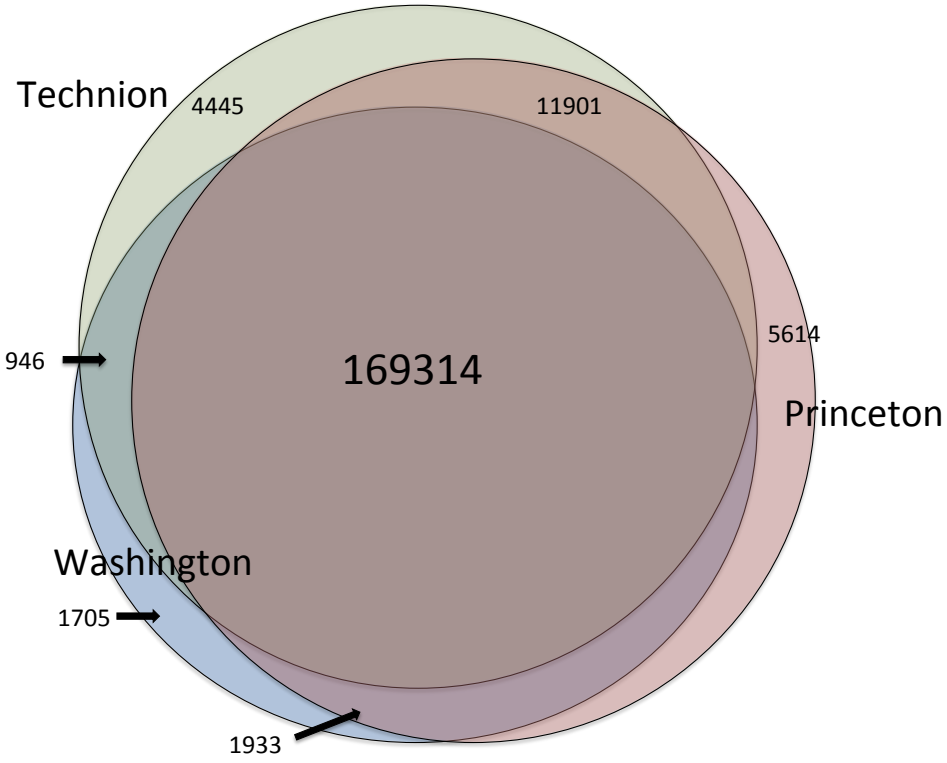


Figure S5 Comparison of SNV calls on various datasets. A Venn diagram shows the overlap of the previous SNV calls (see Table 1) upon comparison with the N2 reference.

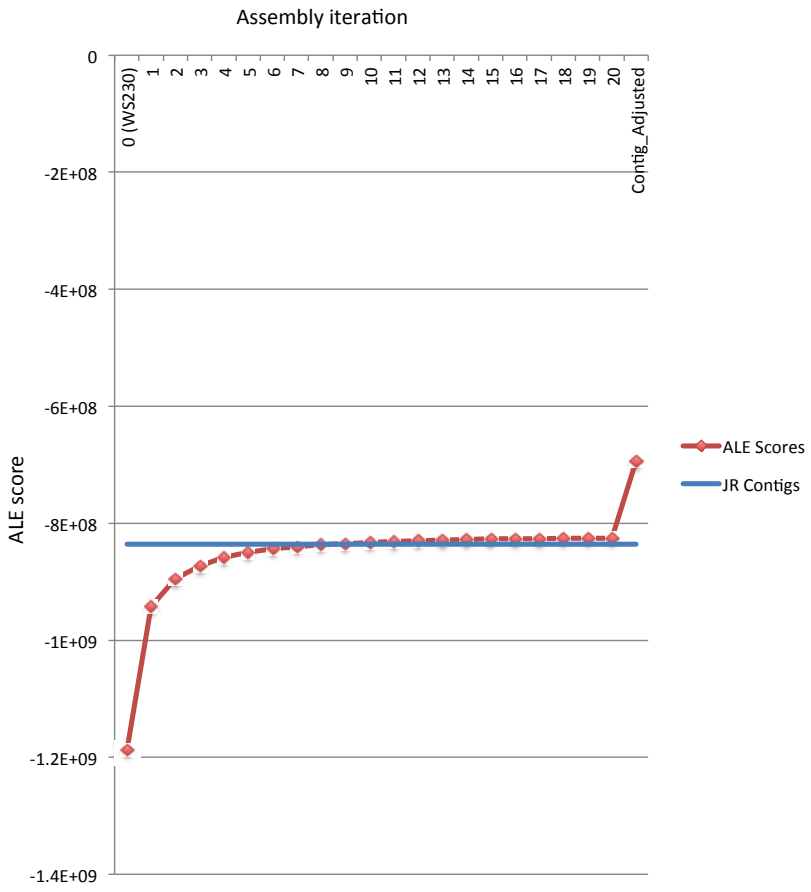


Figure S6 Average ALE scores per base by assembly iteration. The average ALE scores per base, adjusted using N2 reads against the N2 reference as a baseline, plotted for each cycle of the assembly. The average ALE score per base decreased from -6.10 to -2.49 indicating substantial improvement in the quality of the assembly.

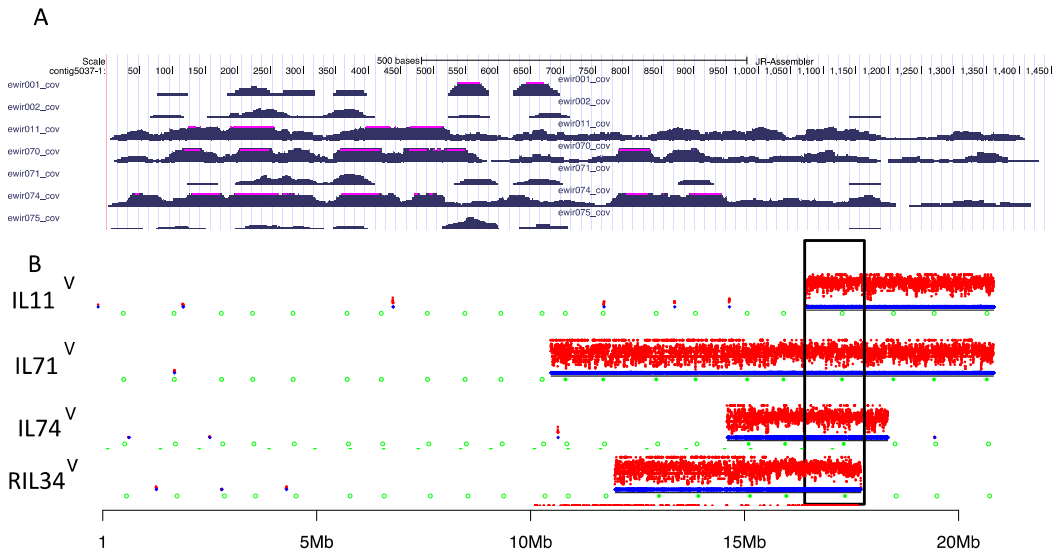
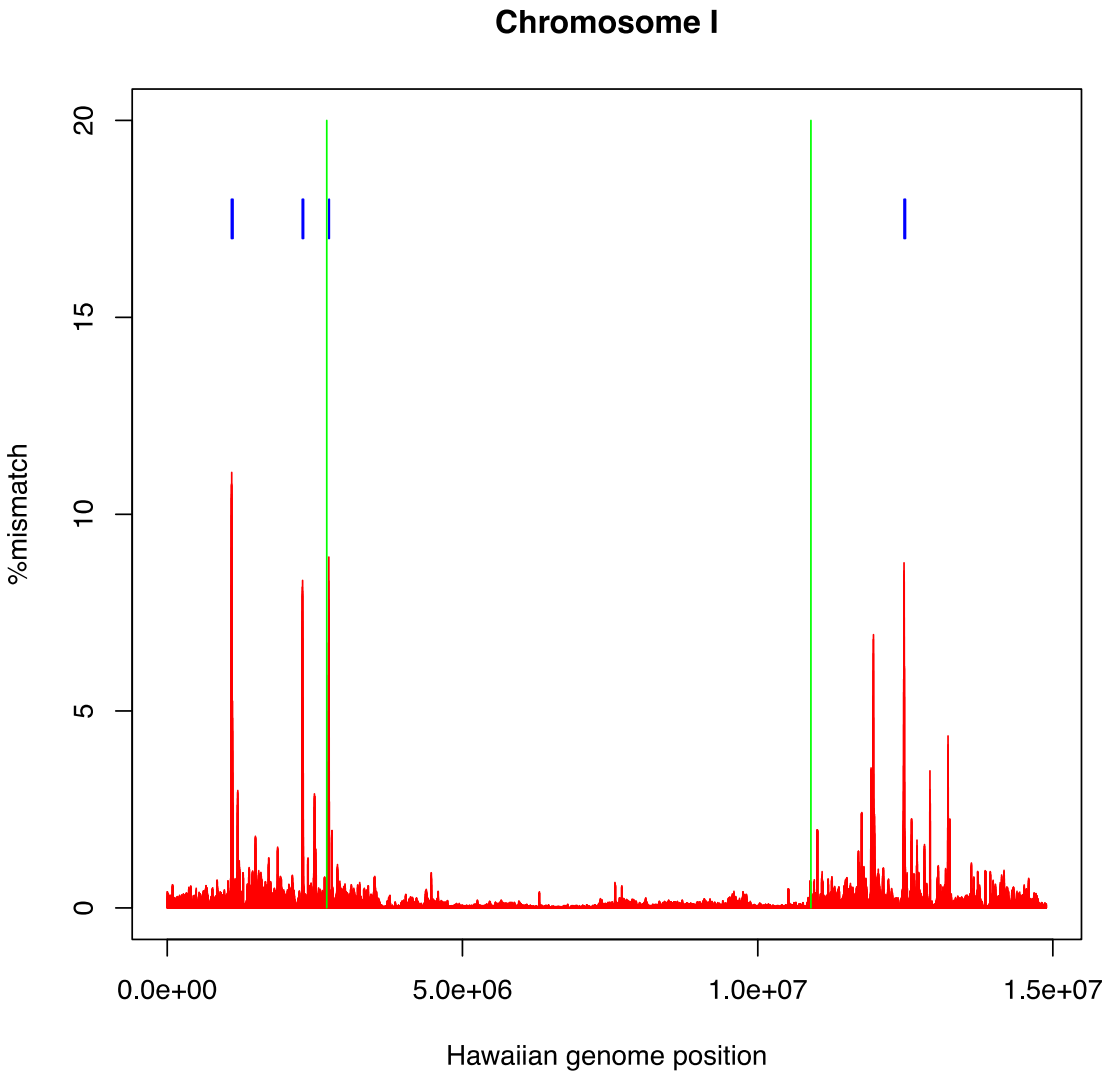
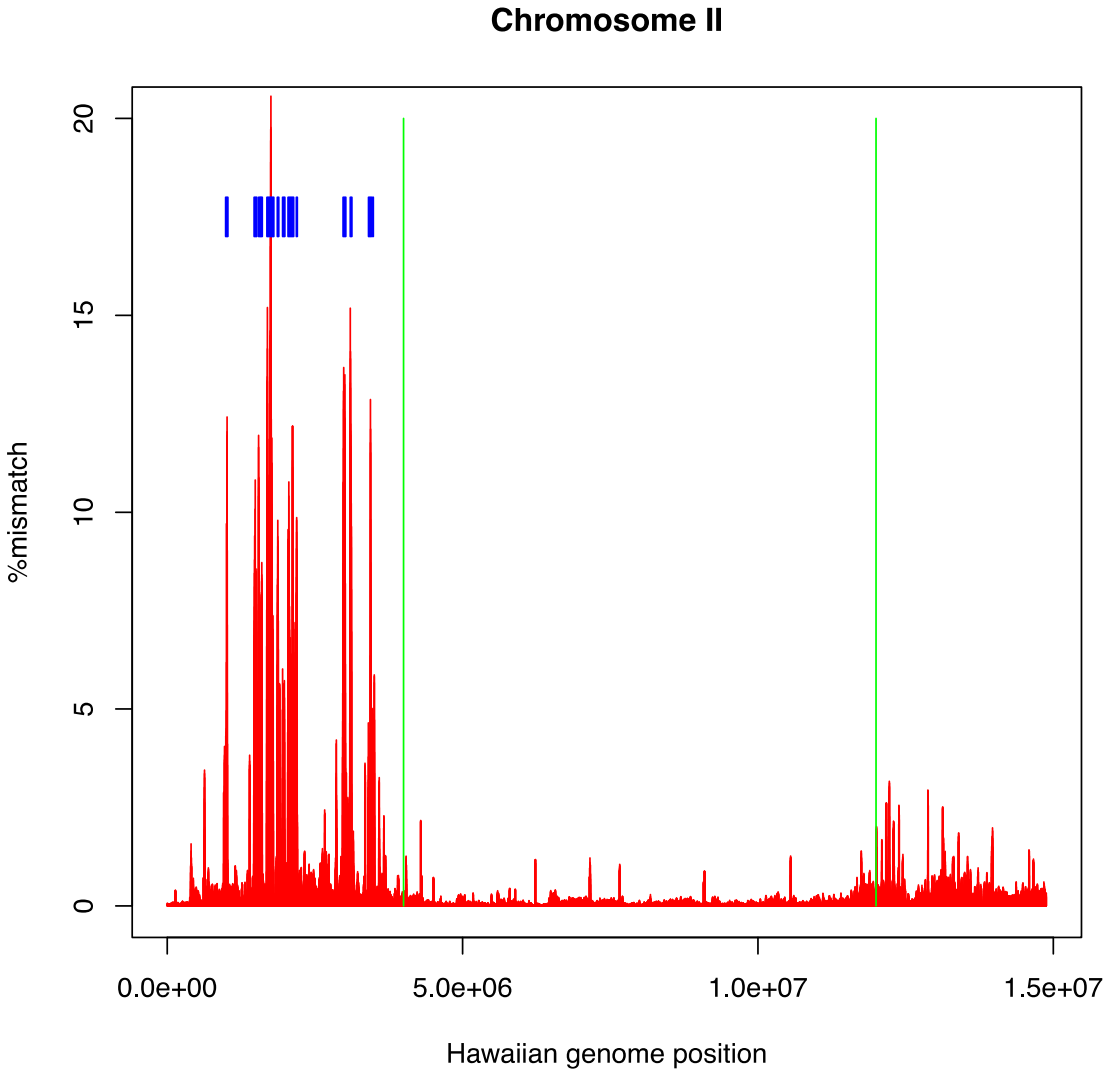
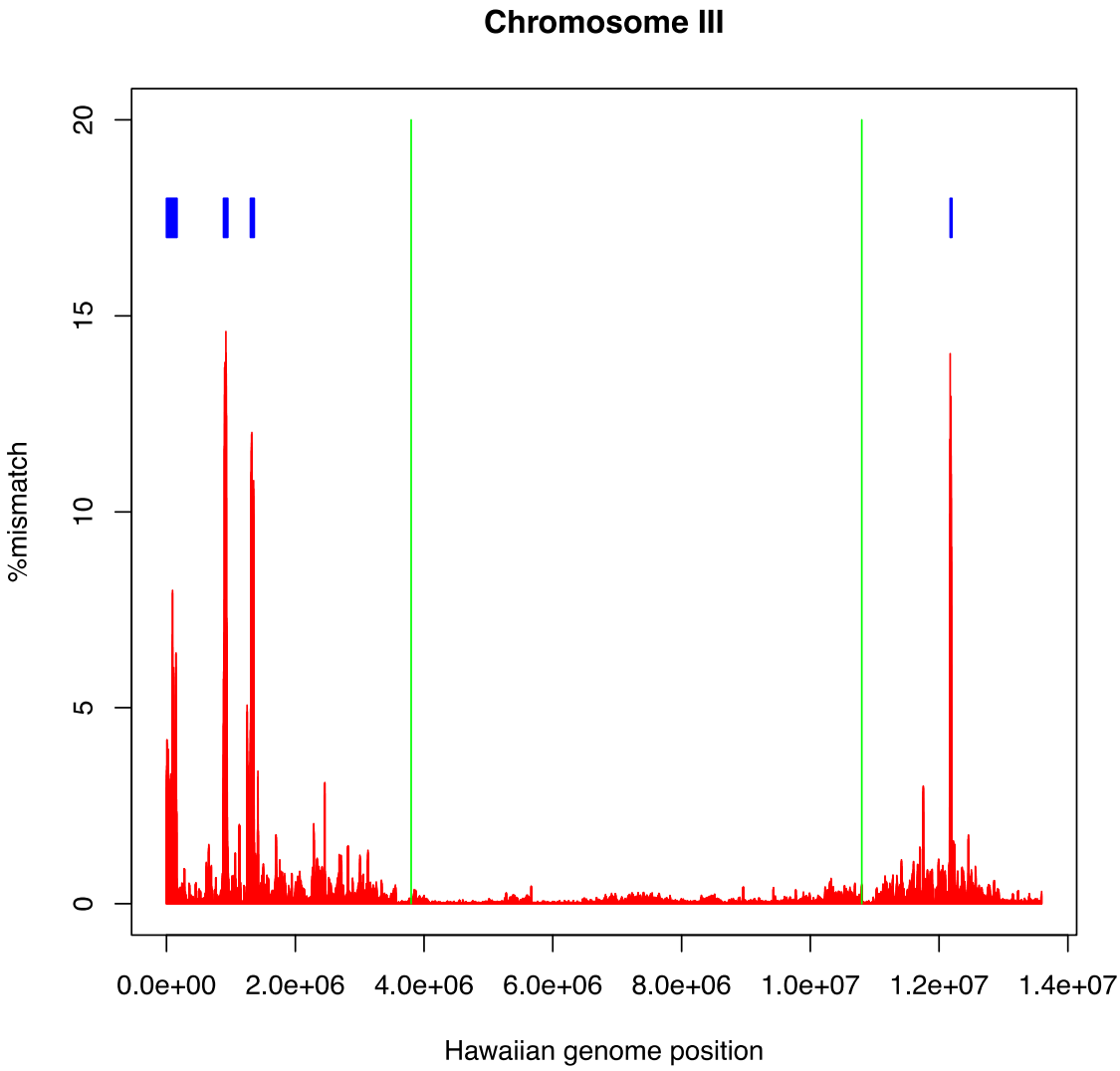
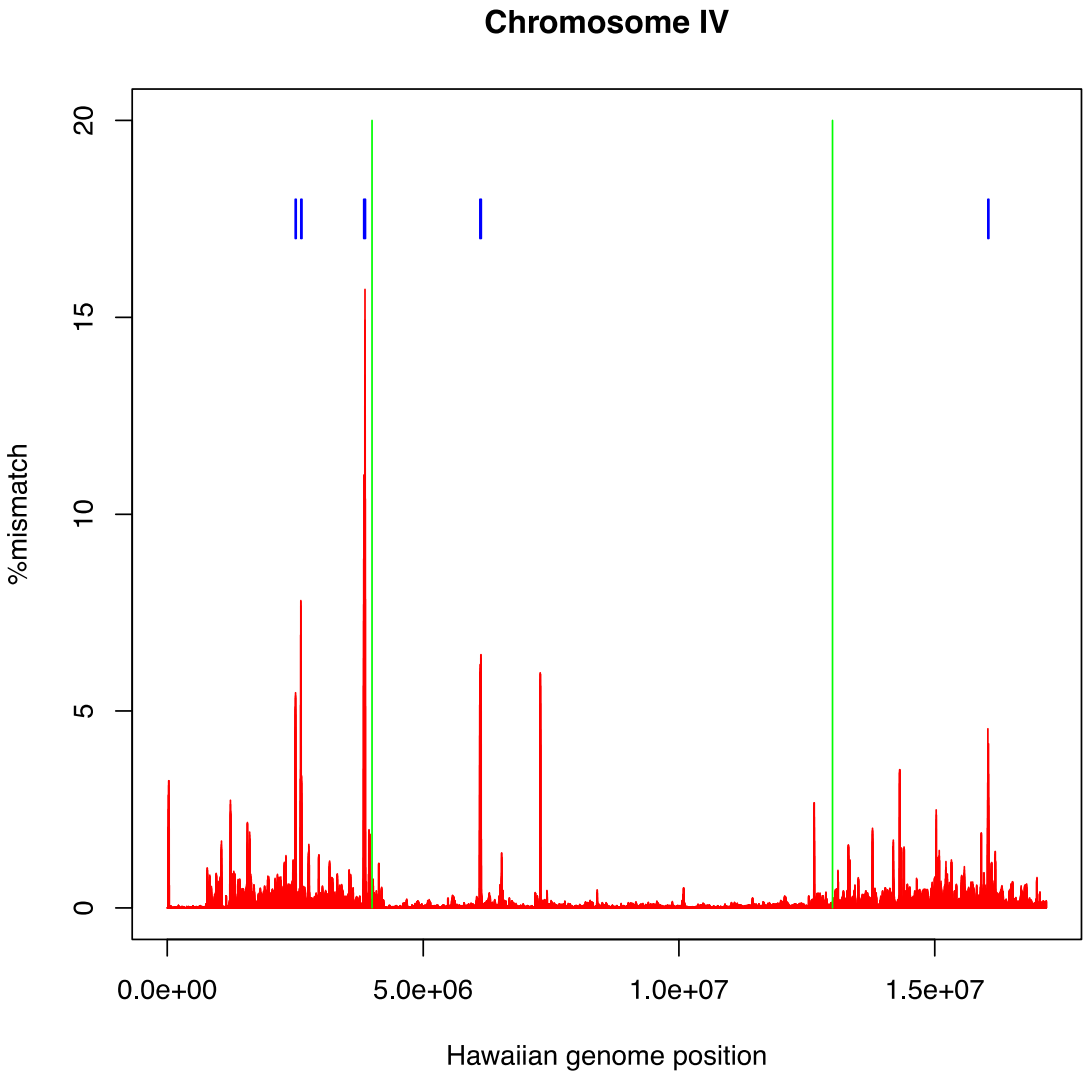


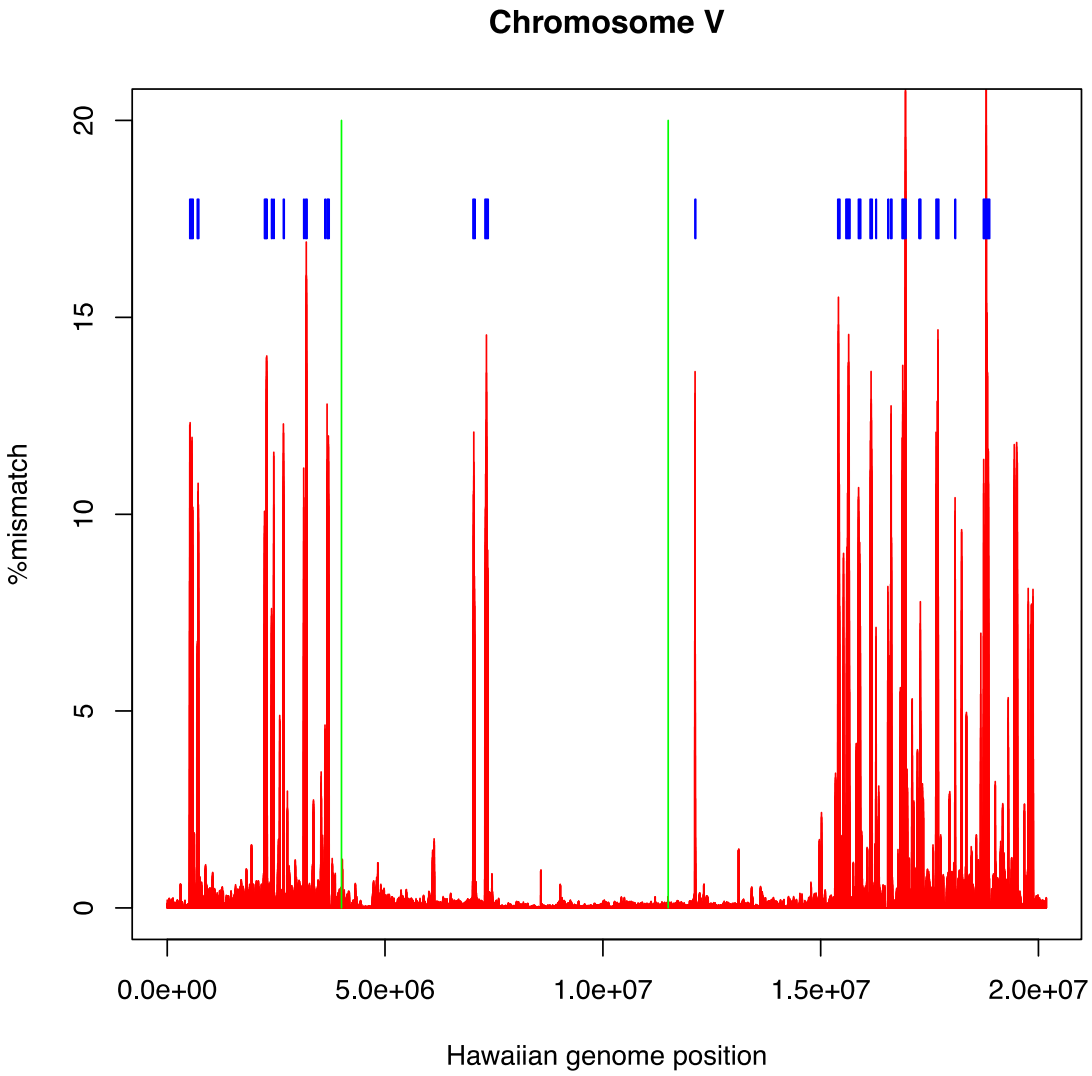
Figure S7 IL read coverage and SNP density. (A) A JR contig (5037-1) with reads aligned against a sample different ILs. A few reads from each of the ILs align against the contig, but only reads from IL 11, 70 and 72 show alignment across most of the length of the contig. This contig must lie in the region of the CB4856 shared by these three lines. (B) Plots of SNV density for chromosome V in the three strains containing CB4856 for JR contig 5037-1. The blocks of red dots indicate regions deriving from CB4856 in these ILs. The CB4856 regions shared between all three strains positions the 5037-1 contig. The RILs could be used to place the cointg still more precisely. The chromosome coordinates in megabases are given below.


Chromosome I
%mismatch
0
5
10
15
20
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Hawaiian genome position


Chromosome II
%mismatch
0
5
10
15
20
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Hawaiian genome position


Chromosome III
%mismatch
0
5
10
15
20
0.0e+00
2.0e+06
4.0e+06
6.0e+06
8.0e+06
1.0e+07
1.2e+07
1.4e+07
Hawaiian genome position


Chromosome IV
%mismatch
0
5
10
15
20
0.0e+00
5.0e+06
1.0e+07
1.5e+07
Hawaiian genome position


Chromosome V
%mismatch
0
5
10
15
20
0.0e+00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Hawaiian genome position

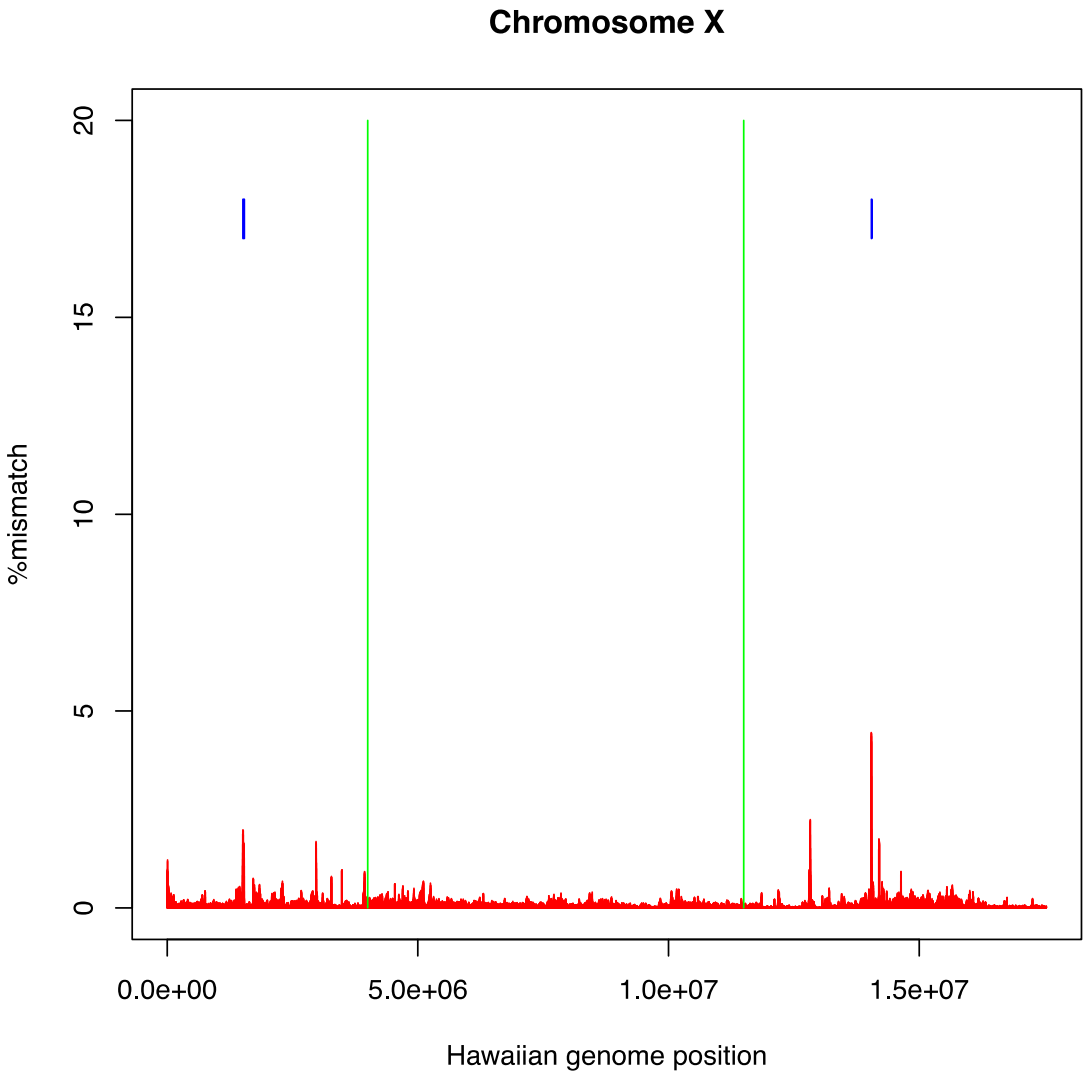


Figure S8 Density of variant sites across chromosomes. Density was calculated in 9kb windows moving in 1kb steps. Blue boxes mark those areas defined as highly divergent. Green lines mark the boundaries of the arms versus centers of the chromosomes.

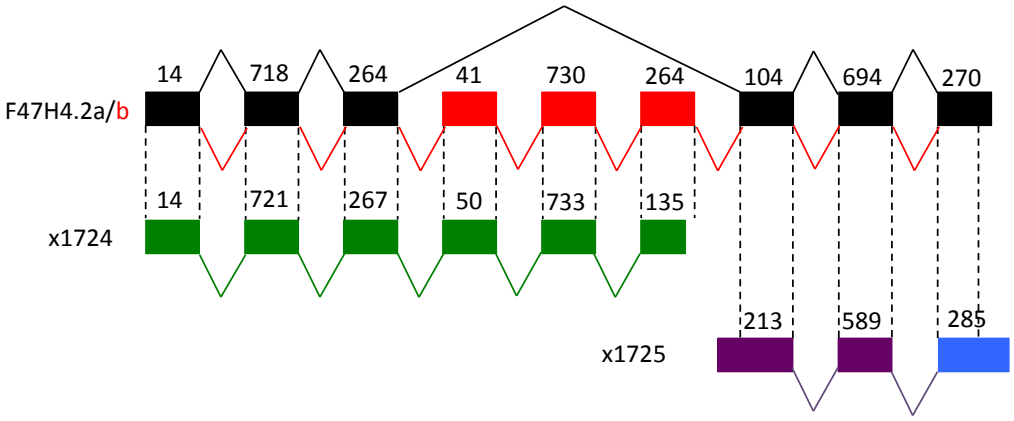
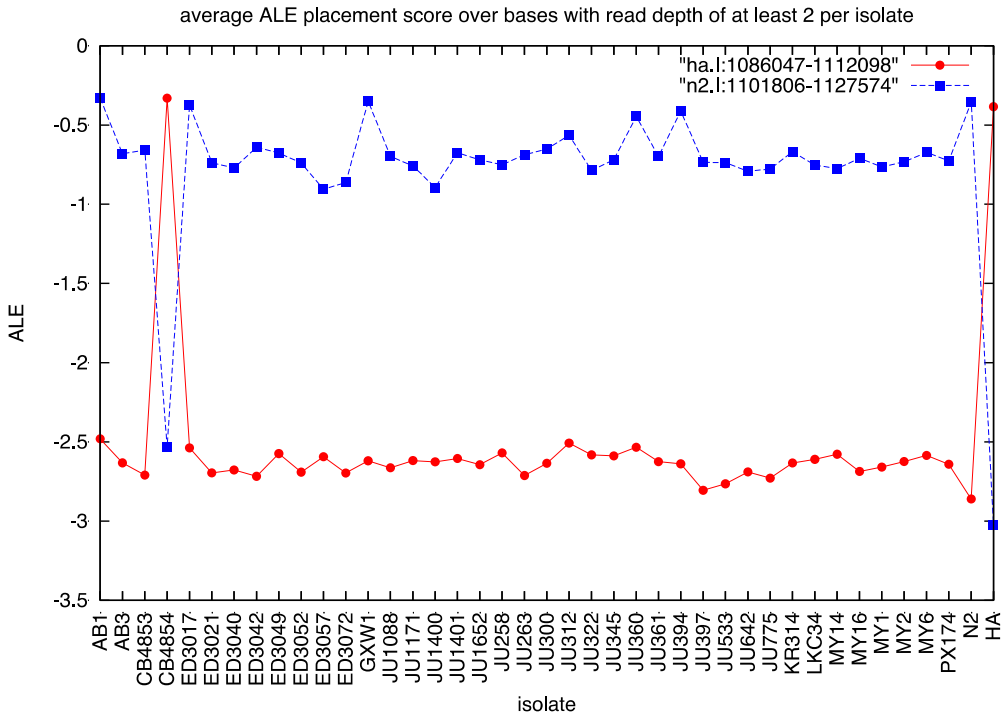
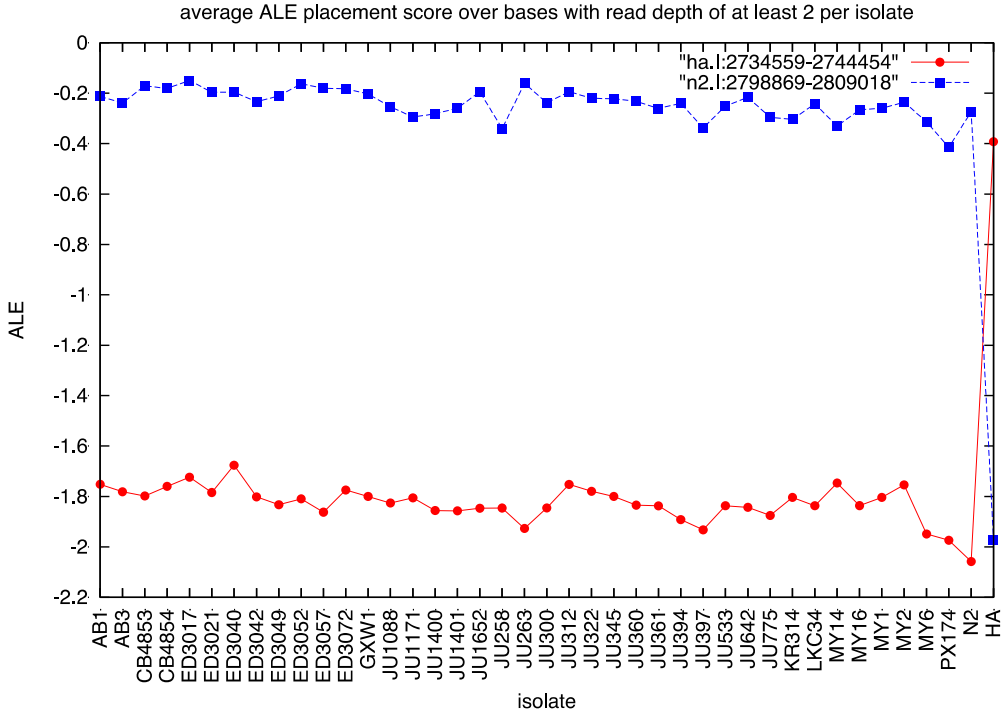
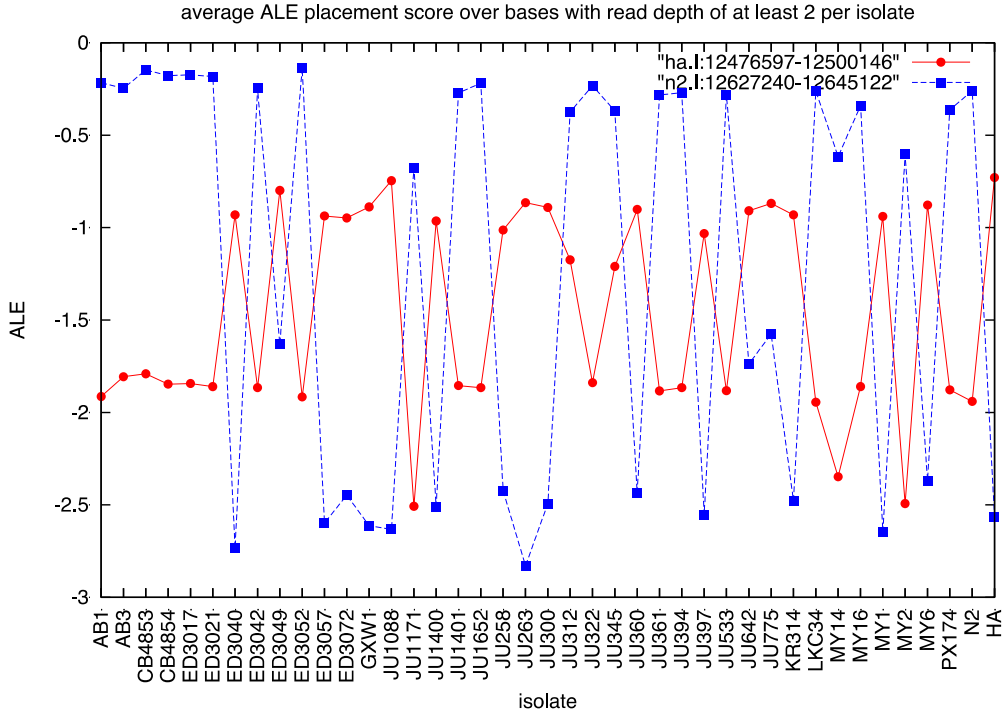
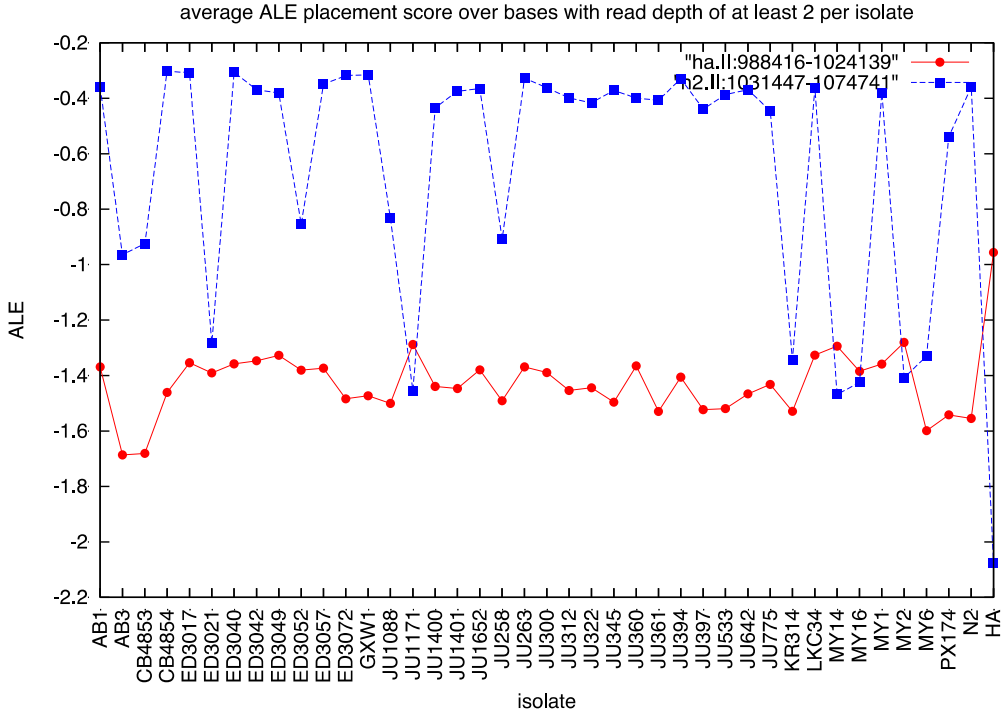


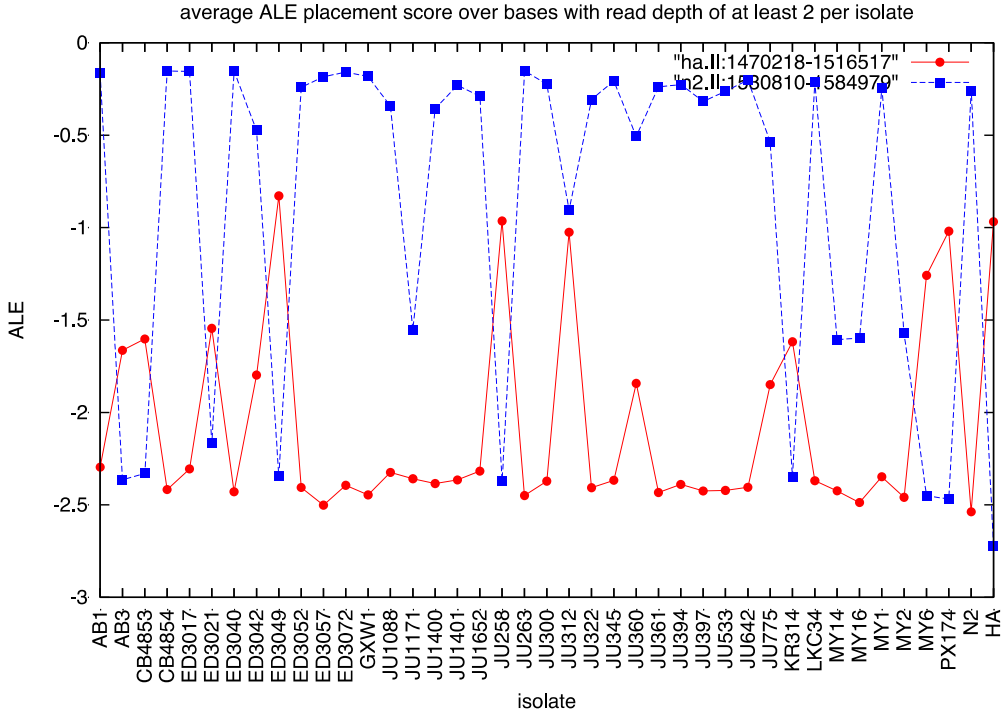
Figure S9 Exon structure of the gene F47H4.2. The exon structure of F47H4.2 (black representing F47H4.2a and red representing F47H4.2b) is shown schematically at the top, with the exon length given above exon box. The black boxes indicate common exons while the red boxes are found in the b isoform only. The exon structure of two Genefinder predictions in the CB4856 reference are shown below. The green boxes indicate the six exons of x1724 and the purple boxes represent two of the exons in prediction x1725. The blue box indicates an open reading frame present in the CB4856 sequence not part of the x1725 model, but in frame with the prior 589 base exon. The dotted lines connect equivalent positions in the two sequences.

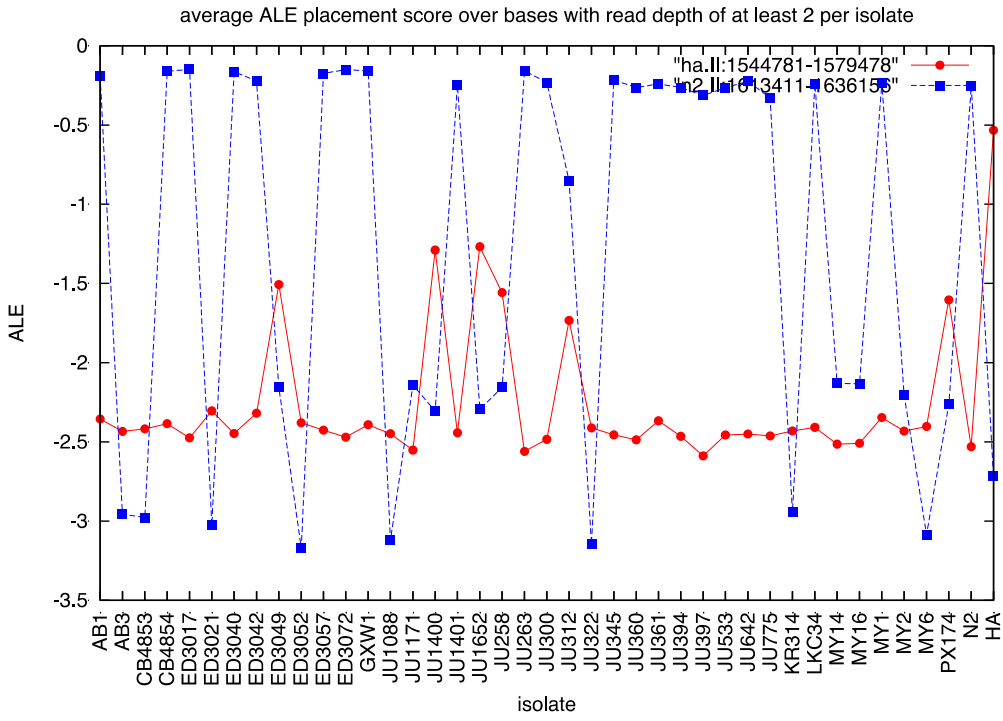

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.I:1086047-1112098"
"n2.I:1101806-1127574"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

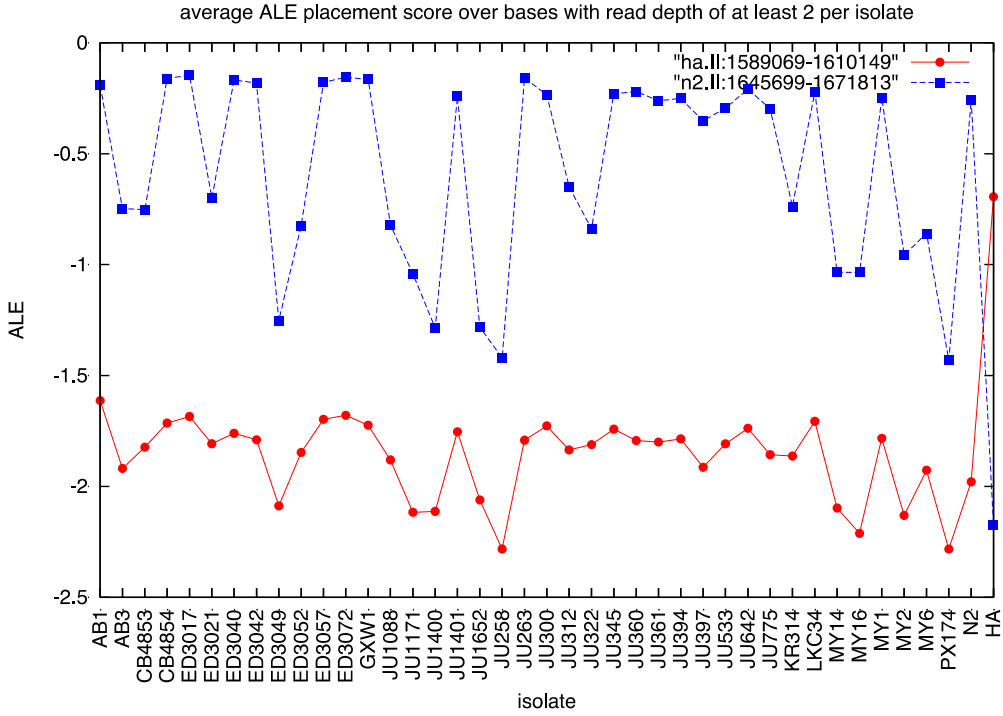

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.I:2734559-2744454"
"n2.I:2798869-2809018"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

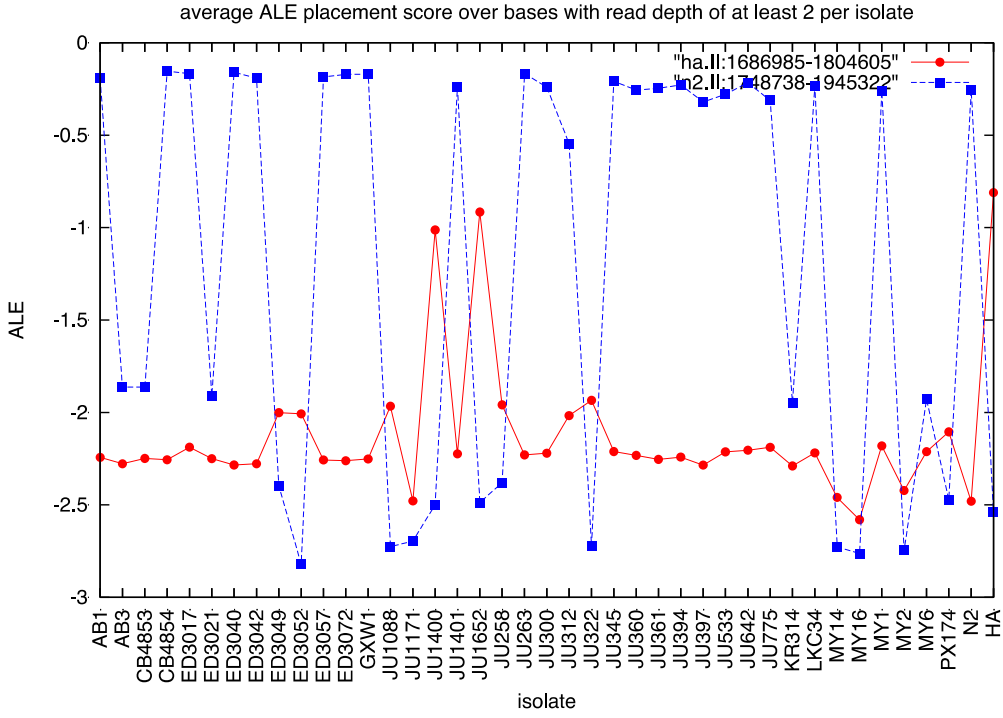

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.I:12476597-12500146"
"n2.I:12627240-12645122"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

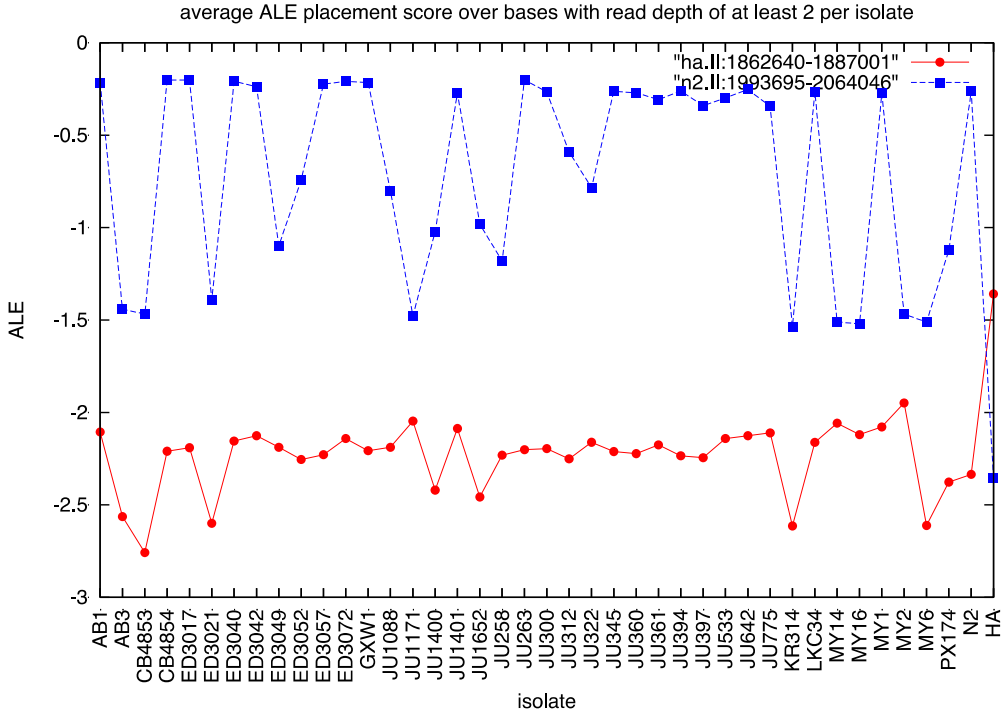

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:988416-1024139"
"h2.II:1031447-1074741"
ALE
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

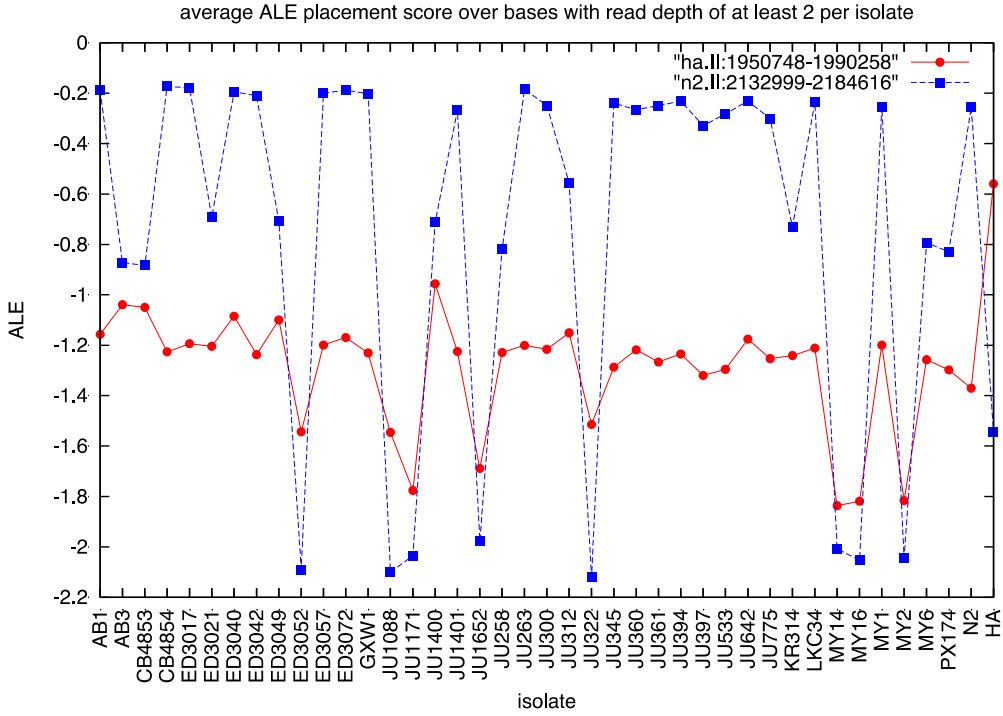

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1470218-1516517"
"m2.II:1530810-1584979"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

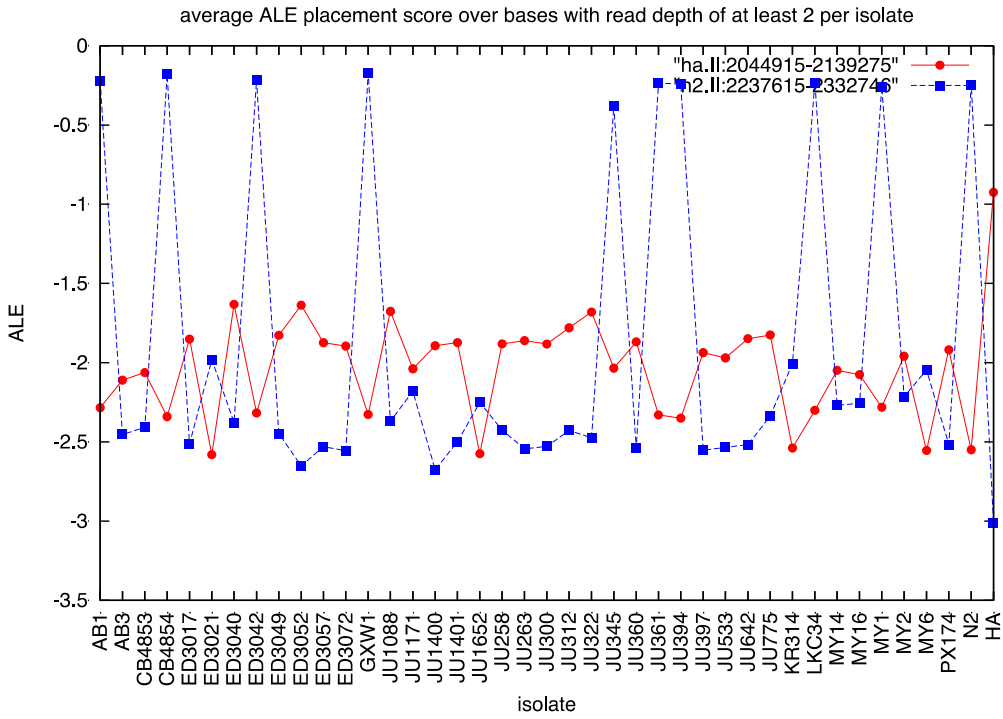

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1544781-1579478"
"n2.II:10113411-10361553"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

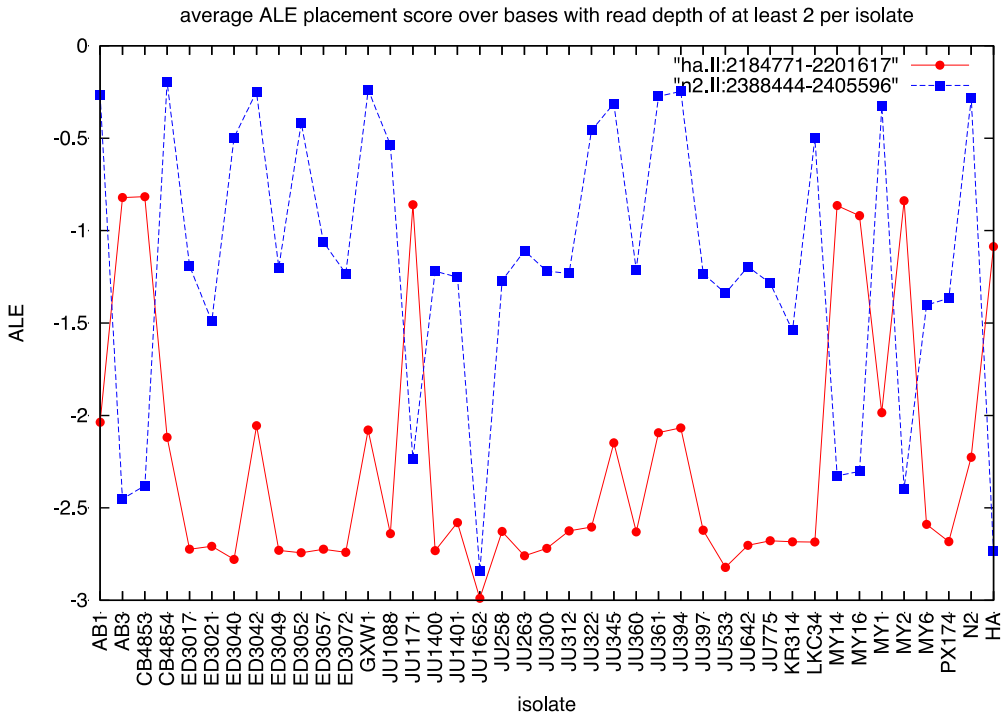

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1589069-1610149"
"n2.II:1645699-1671813"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

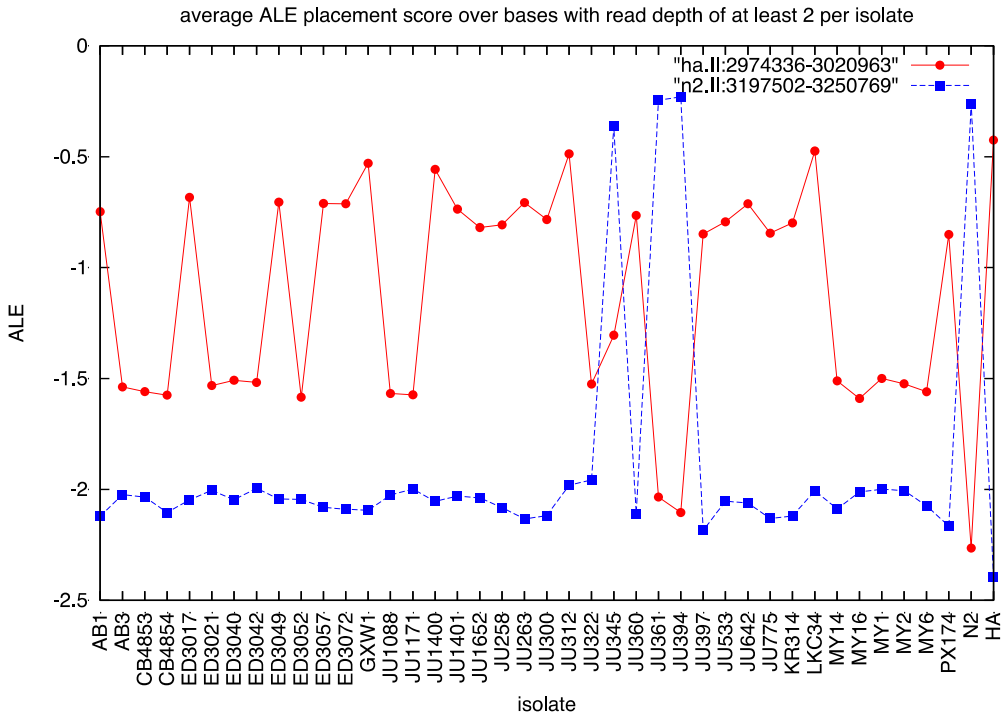

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1686985-1804605"
"n2.II:1748738-1945322"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

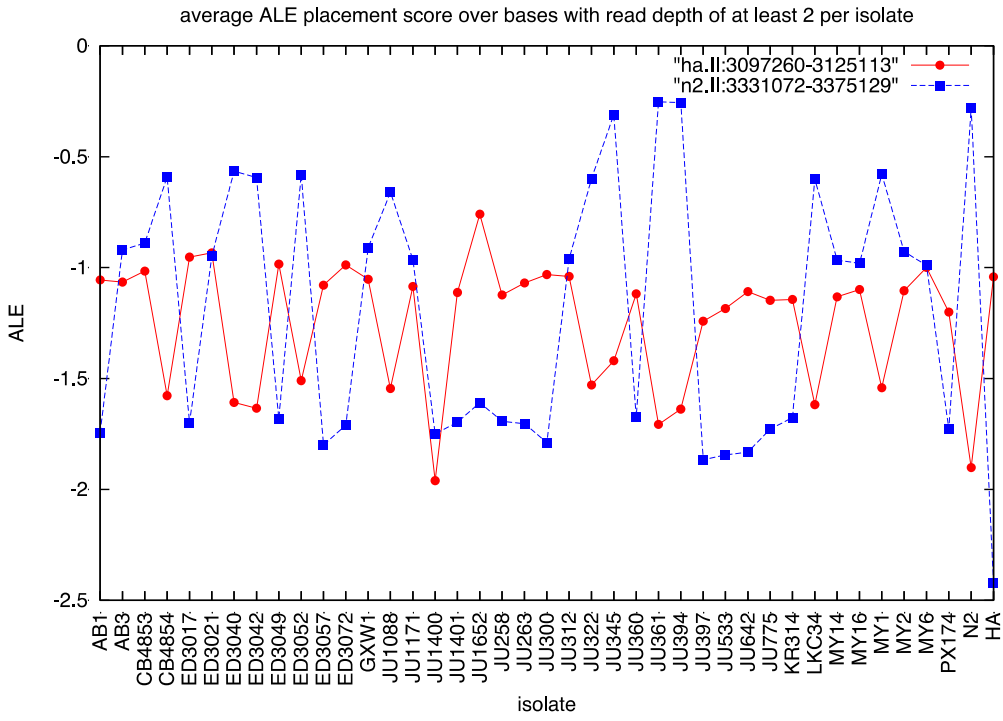

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1862640-1887001"
"n2.II:1993695-2064046"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

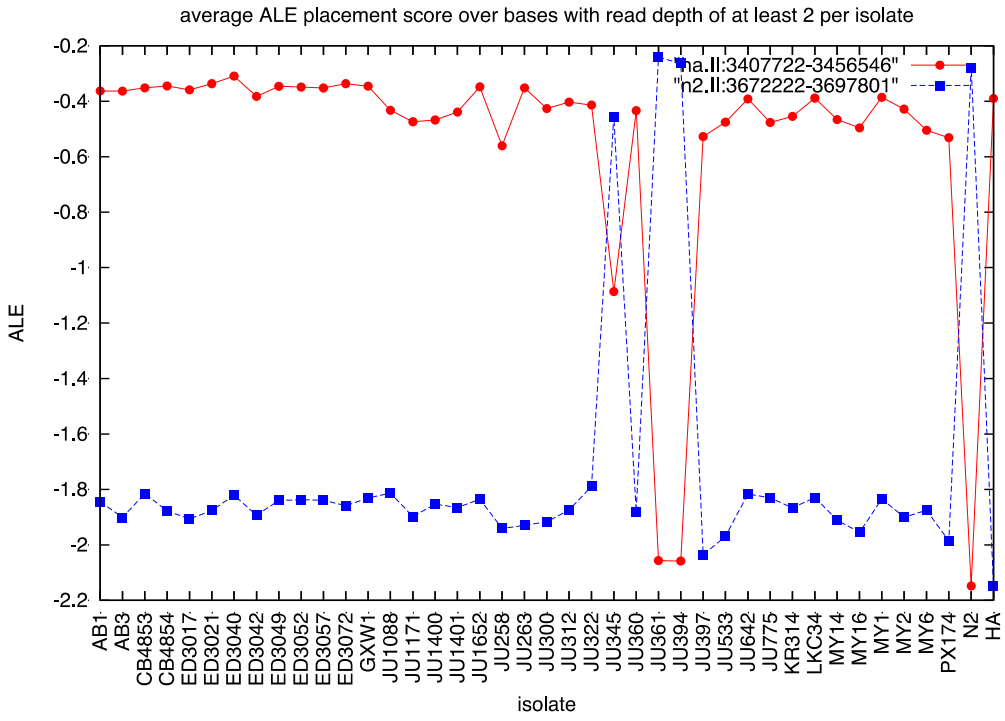

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:1950748-1990258"
"n2.II:2132999-2184616"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

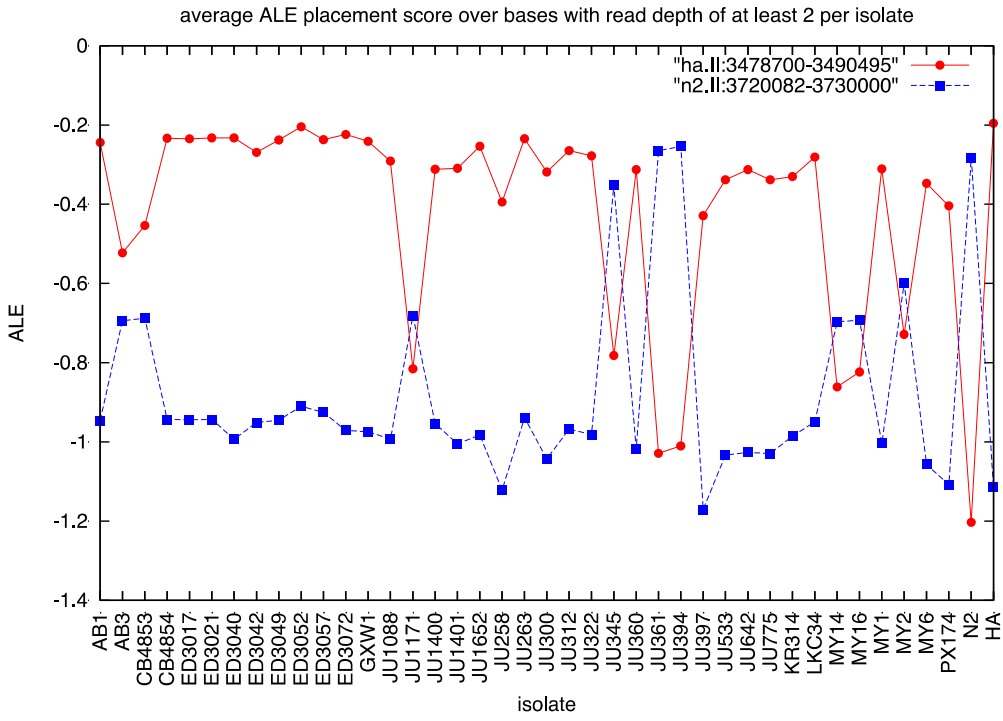

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:2044915-2139275"
"n2.II:2237615-2332746"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

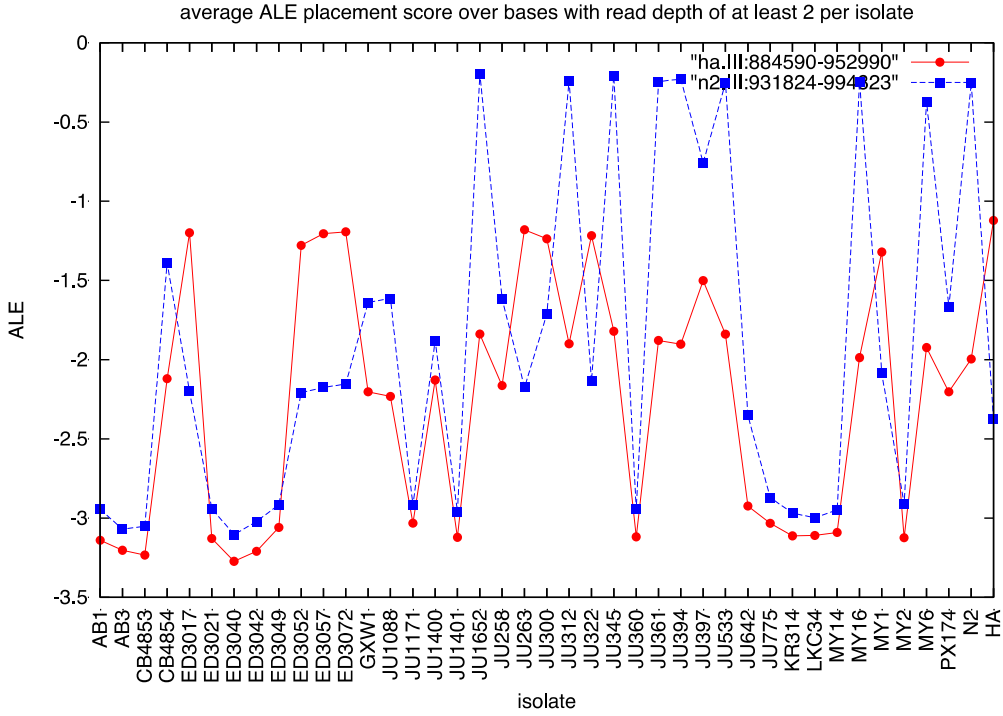

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:2184771-2201617"
"n2.II:2388444-2405596"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

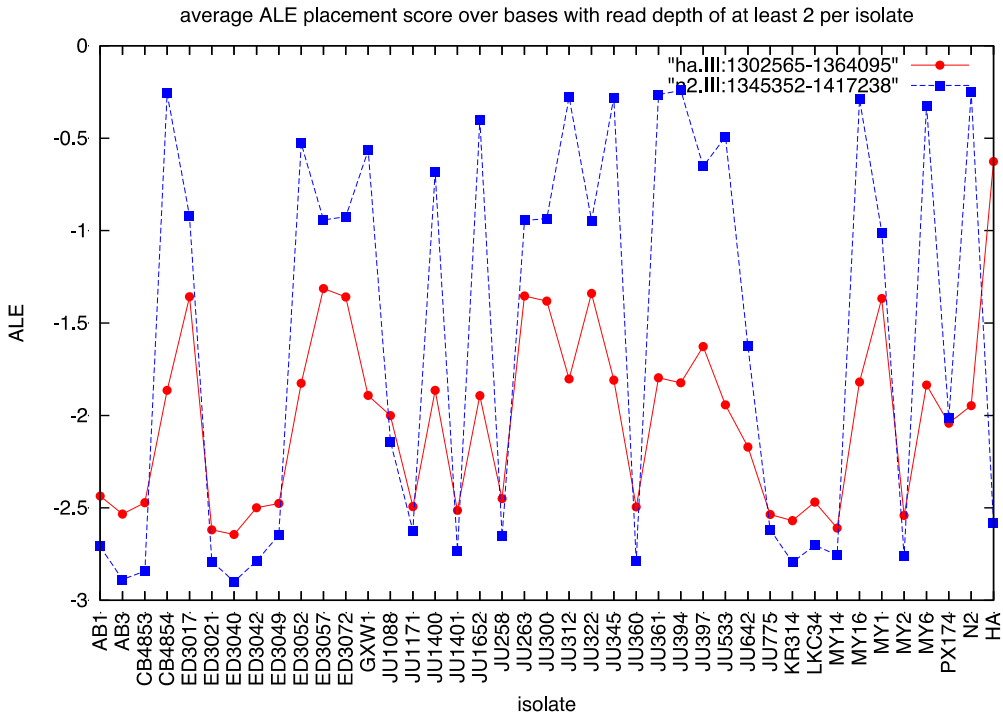

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:2974336-3020963"
"n2.II:3197502-3250769"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

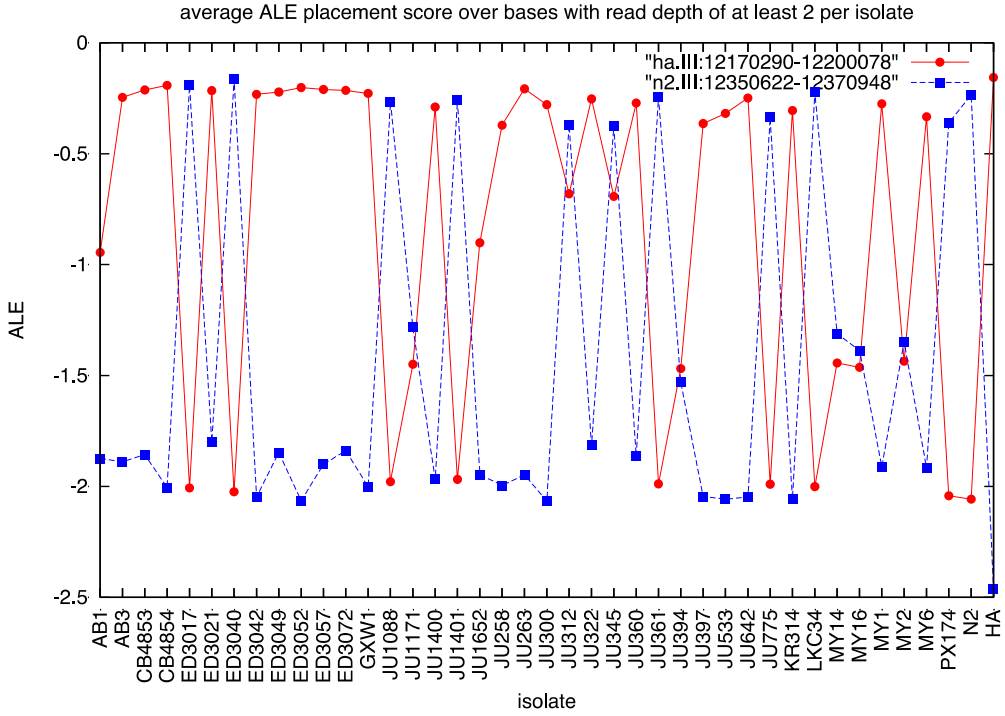

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:3097260-3125113"
"n2.II:3331072-3375129"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

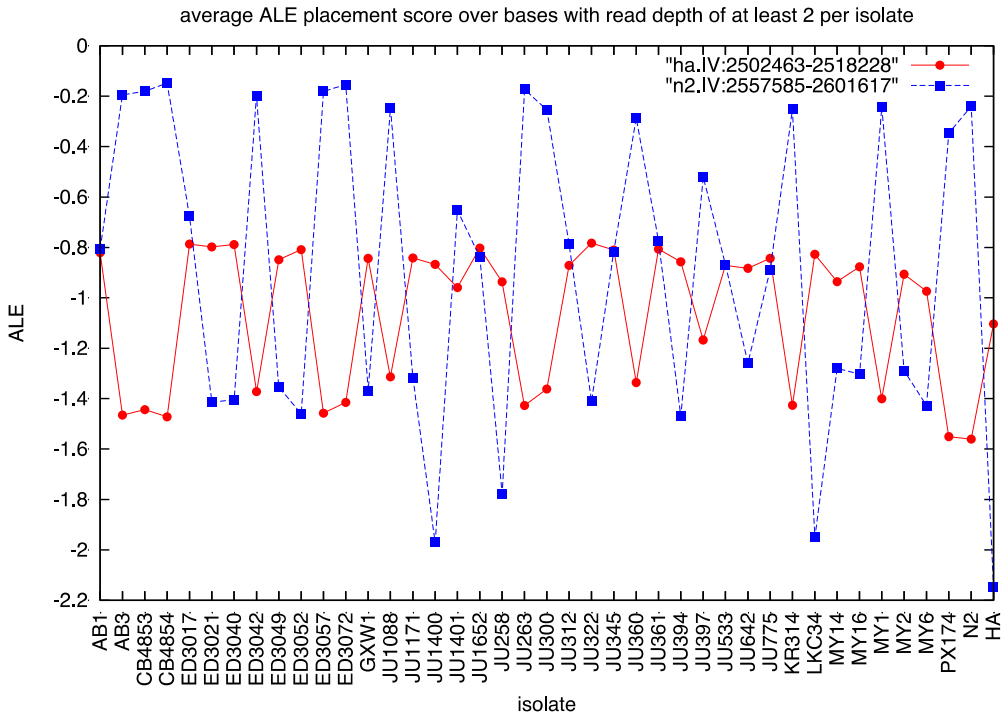

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:3407722-3456546"
"h2.II:3672222-3697801"
ALE
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

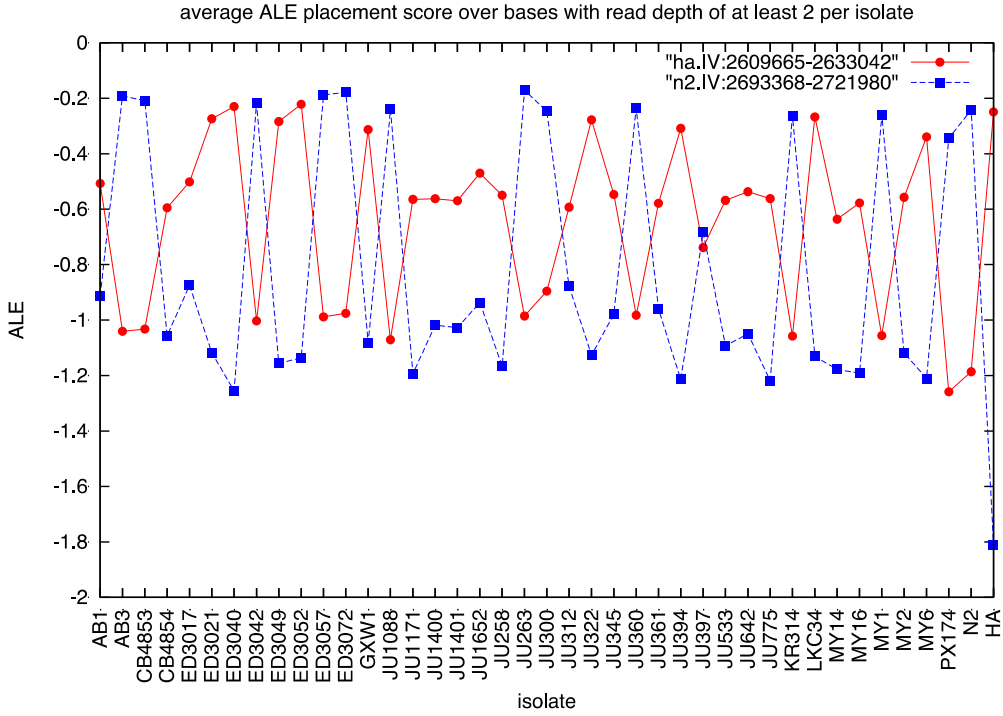

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.II:3478700-3490495"
"n2.II:3720082-3730000"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

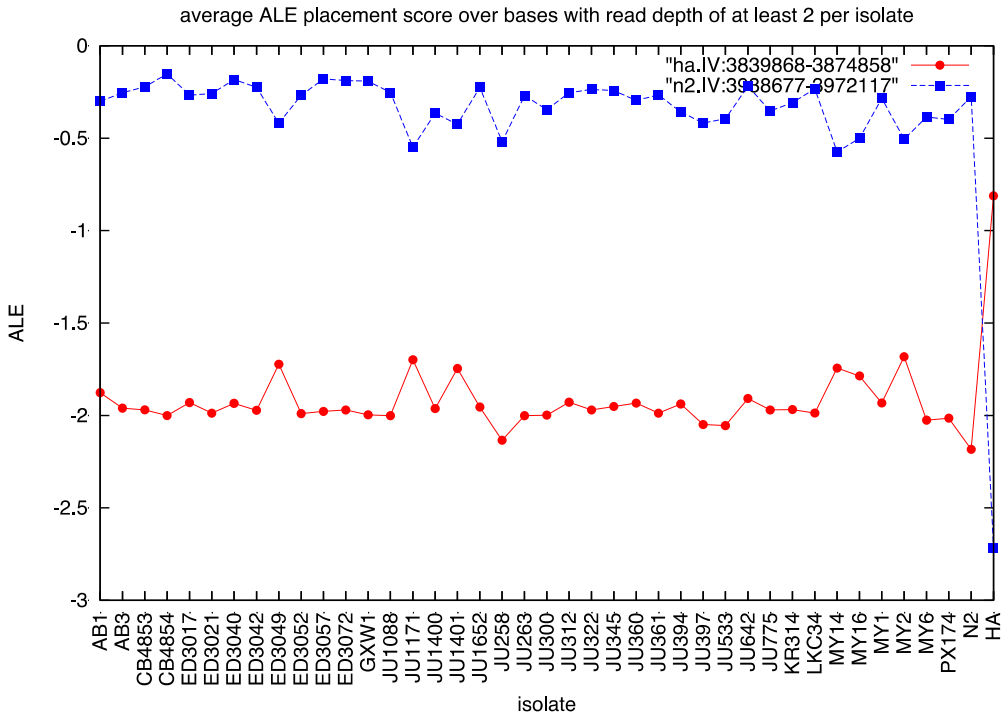

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.III:884590-952990"
"n2.III:931824-994323"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

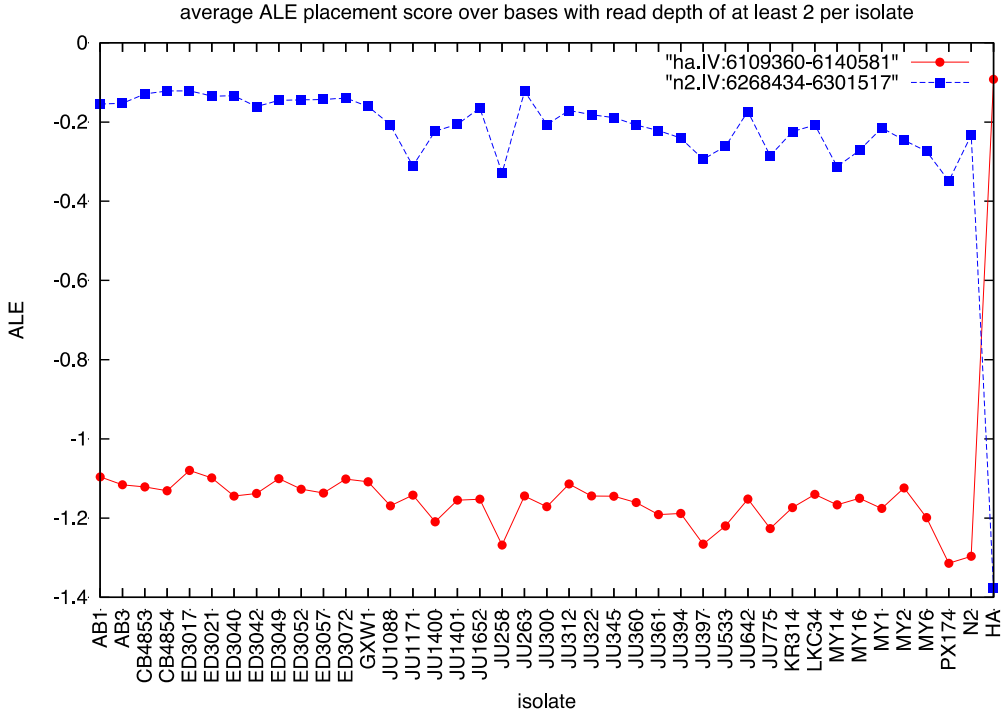

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.III:1302565-1364095"
"n2.III:1345352-1417238"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

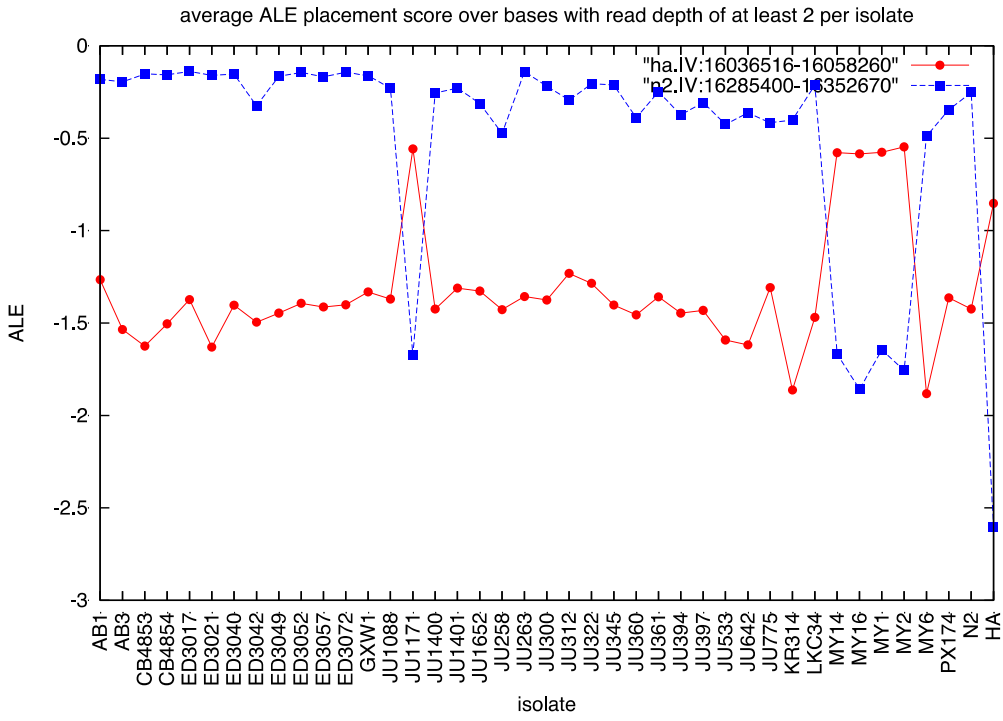

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.III:12170290-12200078"
"n2.III:12350622-12370948"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

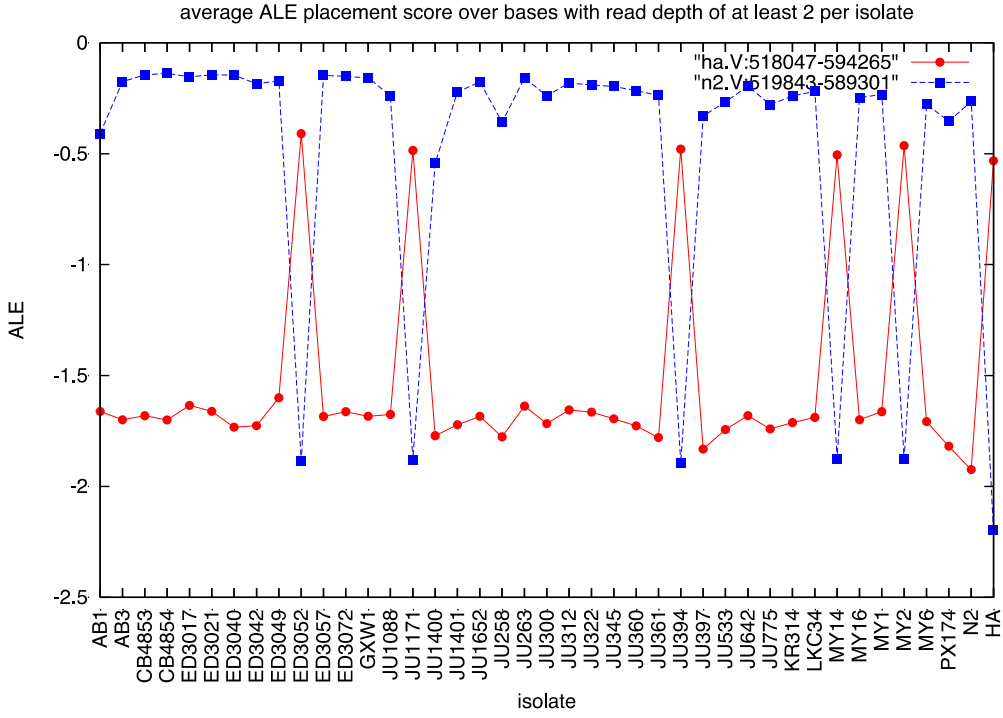

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.IV:2502463-2518228"
"n2.IV:2557585-2601617"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

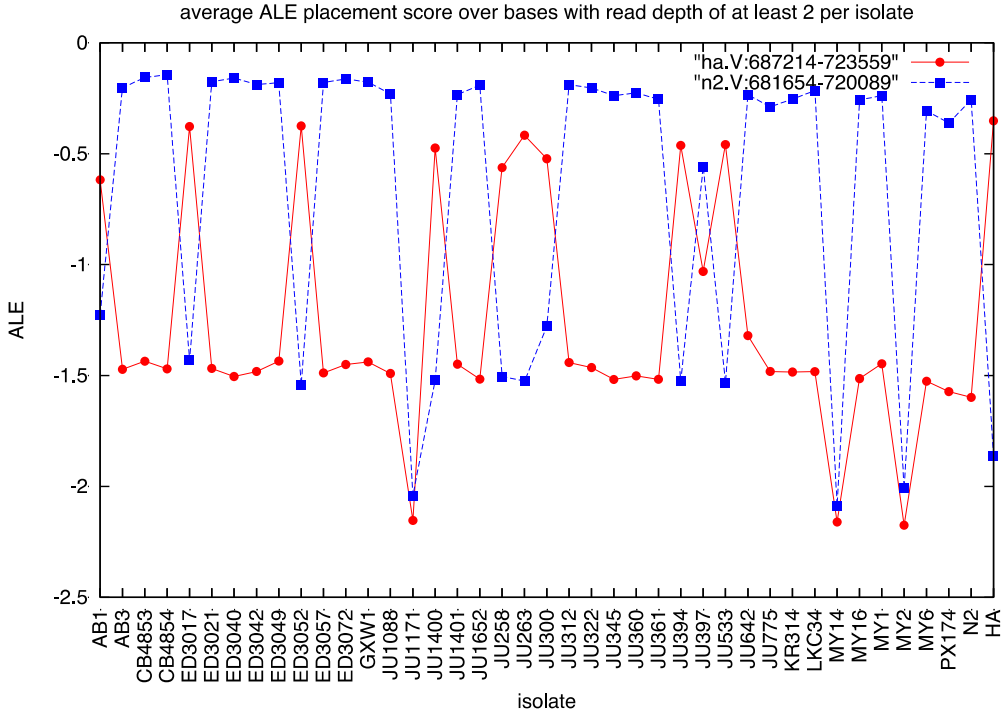

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.IV:2609665-2633042"
"n2.IV:2693368-2721980"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

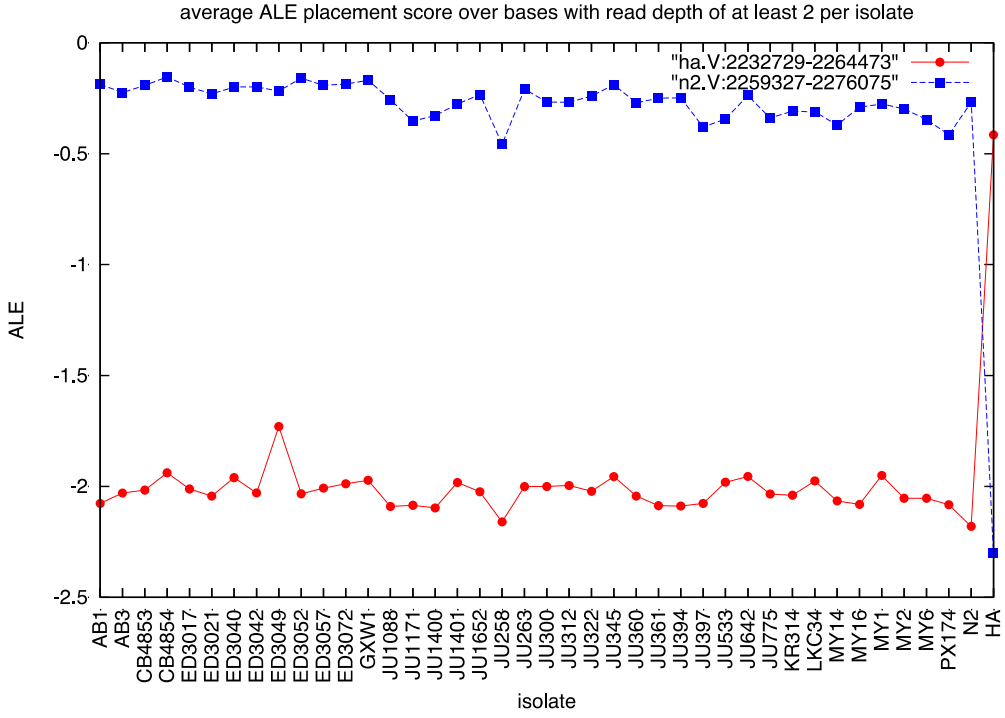

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.IV:3839868-3874858"
"n2.IV:3988677-3972117"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

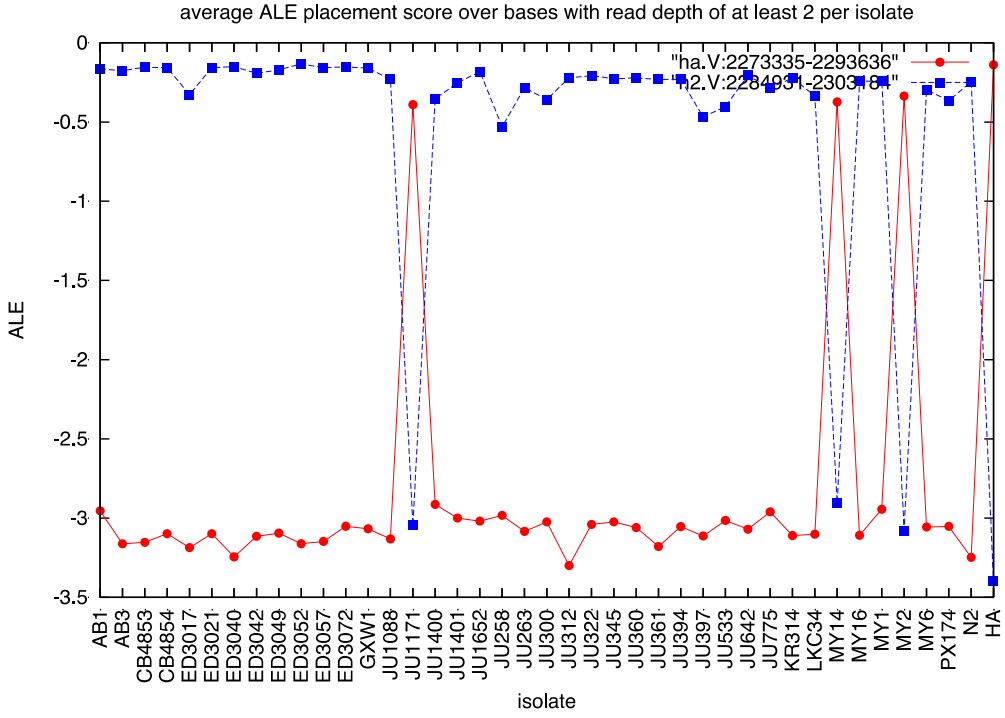

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.IV:6109360-6140581"
"n2.IV:6268434-6301517"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

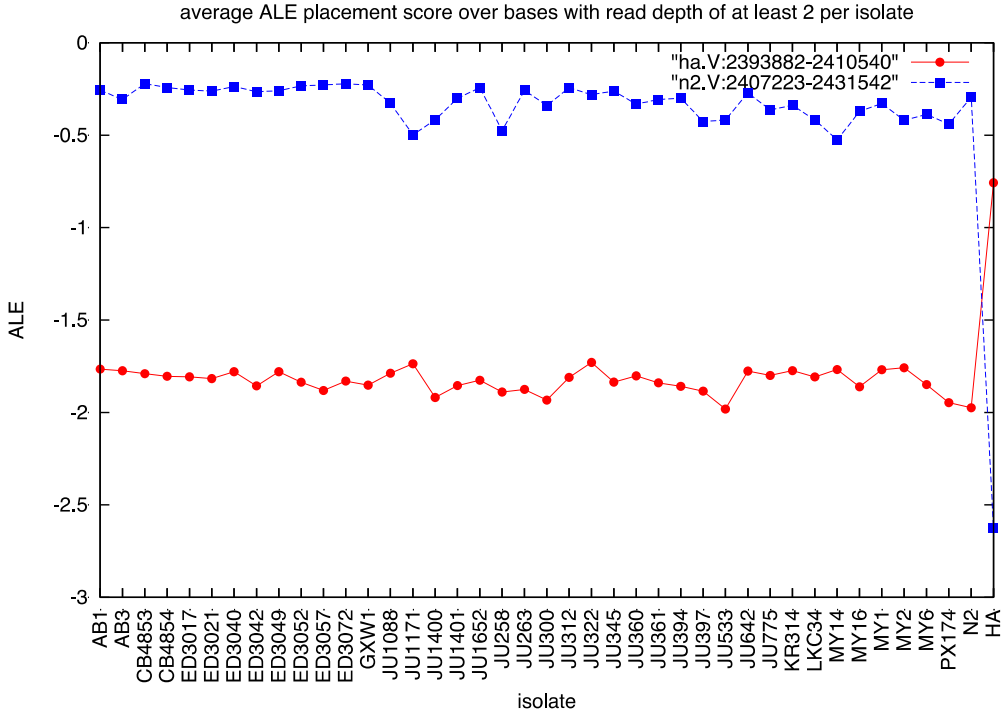

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.IV:16036516-16058260"
"n2.IV:16285400-16352670"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

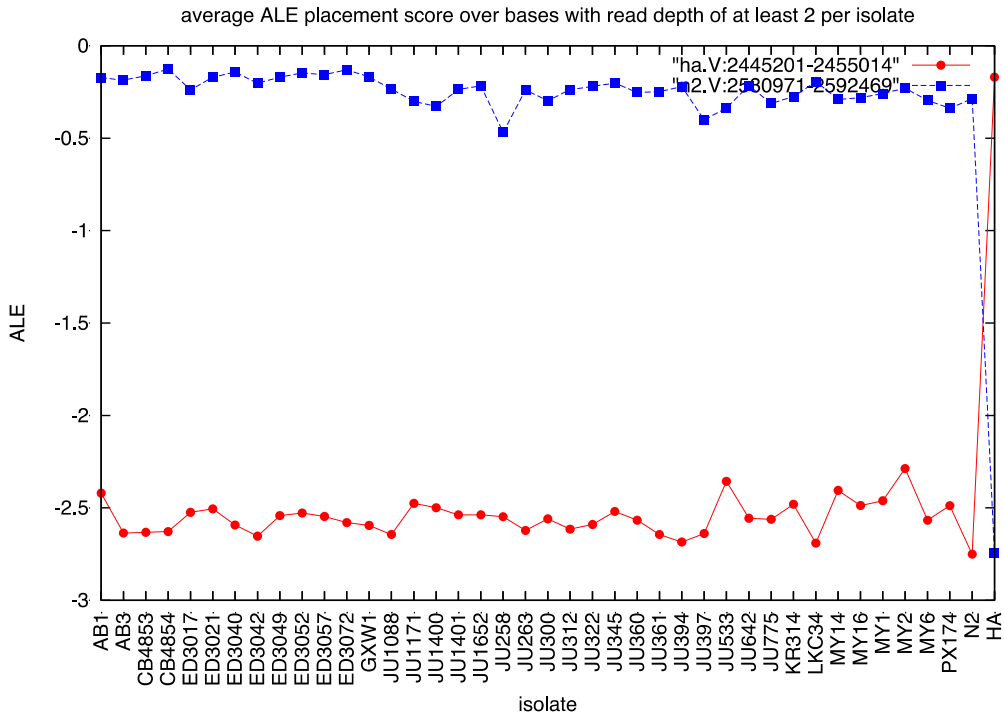

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:518047-594265"
"n2.V:519843-589301"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

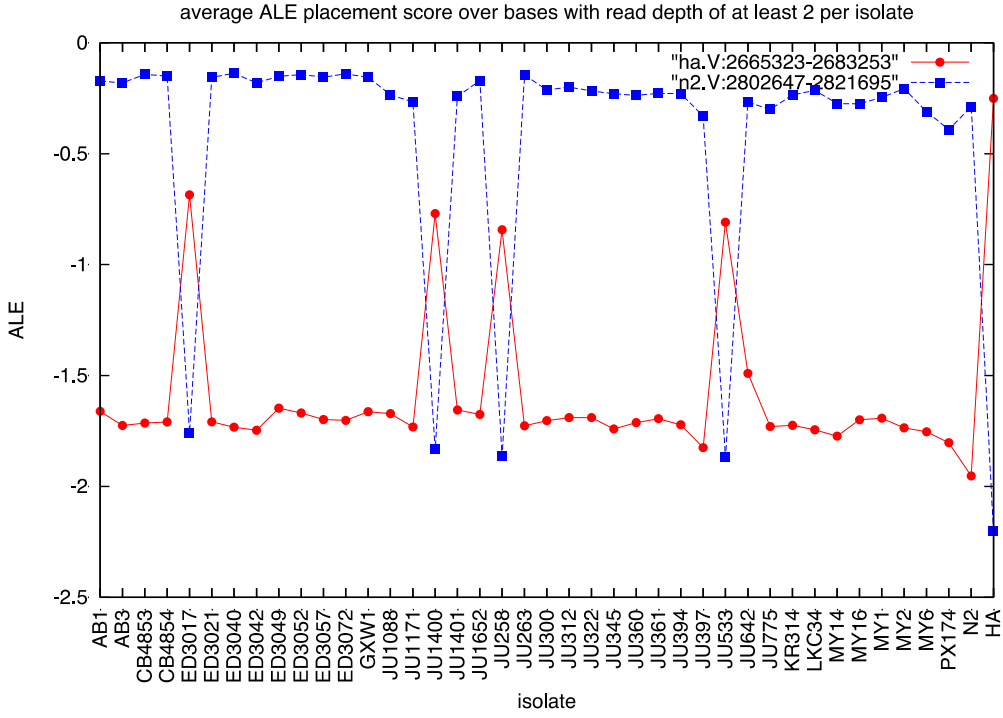

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:687214-723559"
"n2.V:681654-720089"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

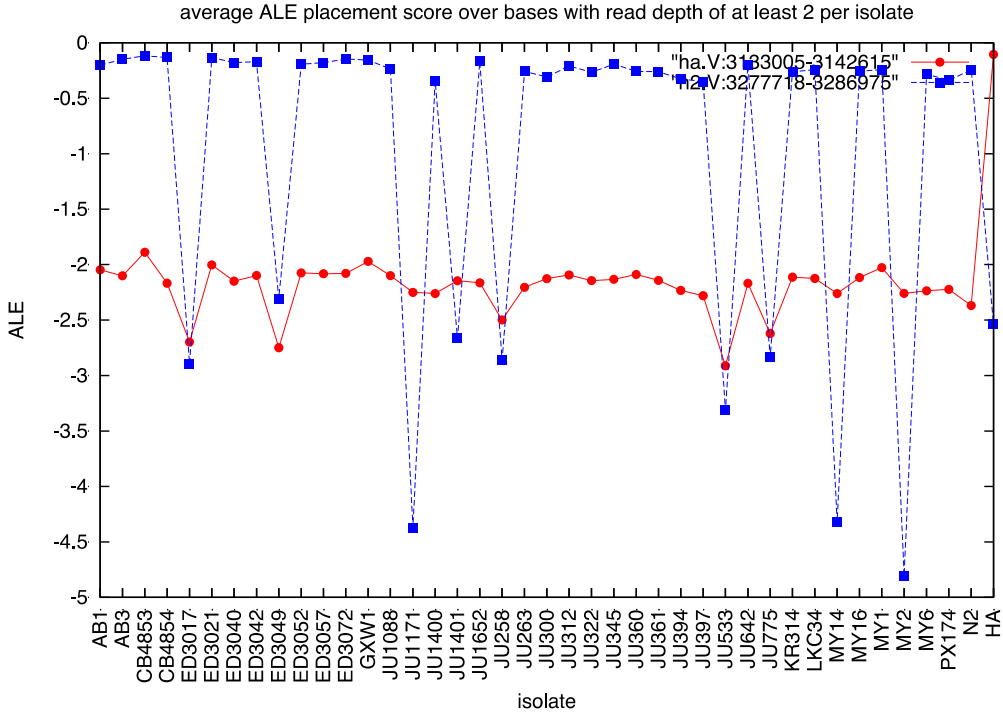

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:2232729-2264473"
"n2.V:2259327-2276075"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

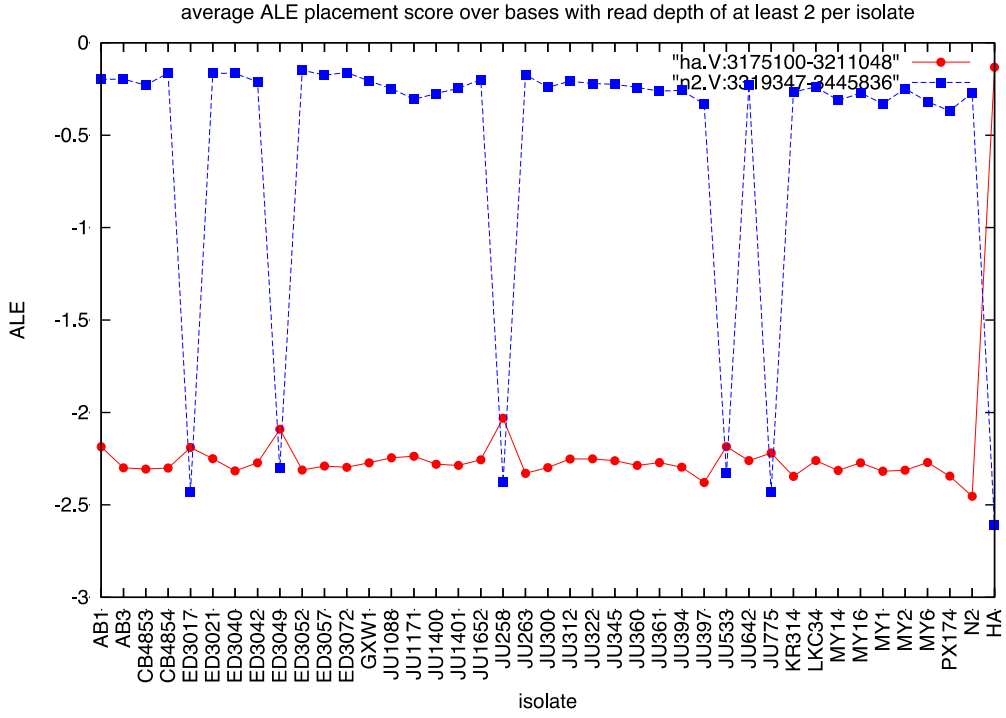

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:2273335-2293636"
"h2.V:2284931-2303184"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

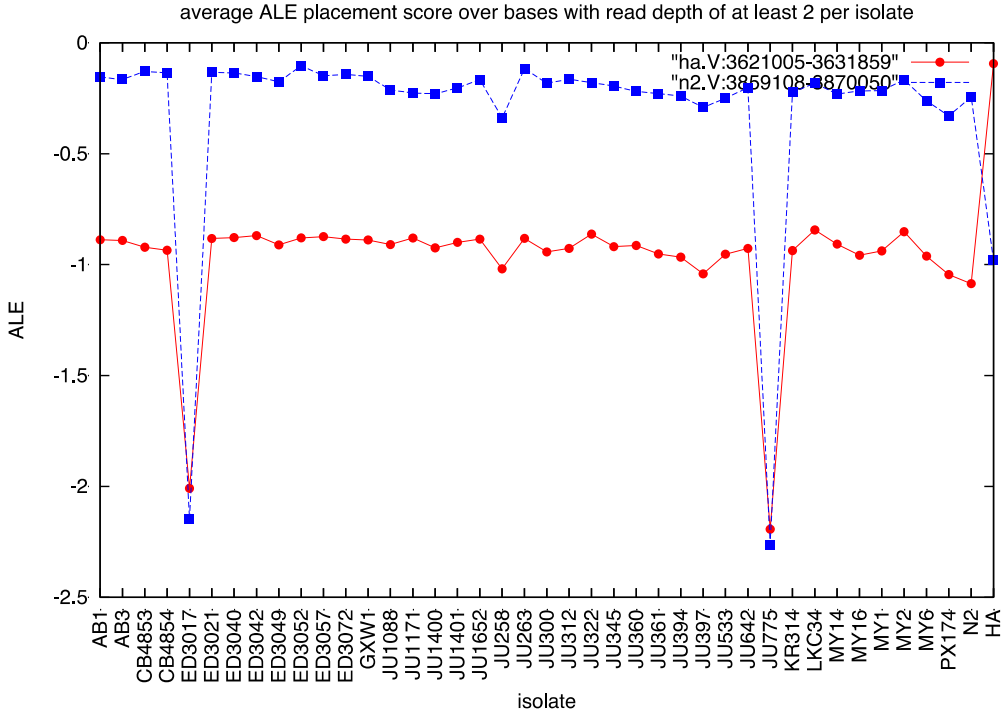

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:2393882-2410540"
"n2.V:2407223-2431542"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

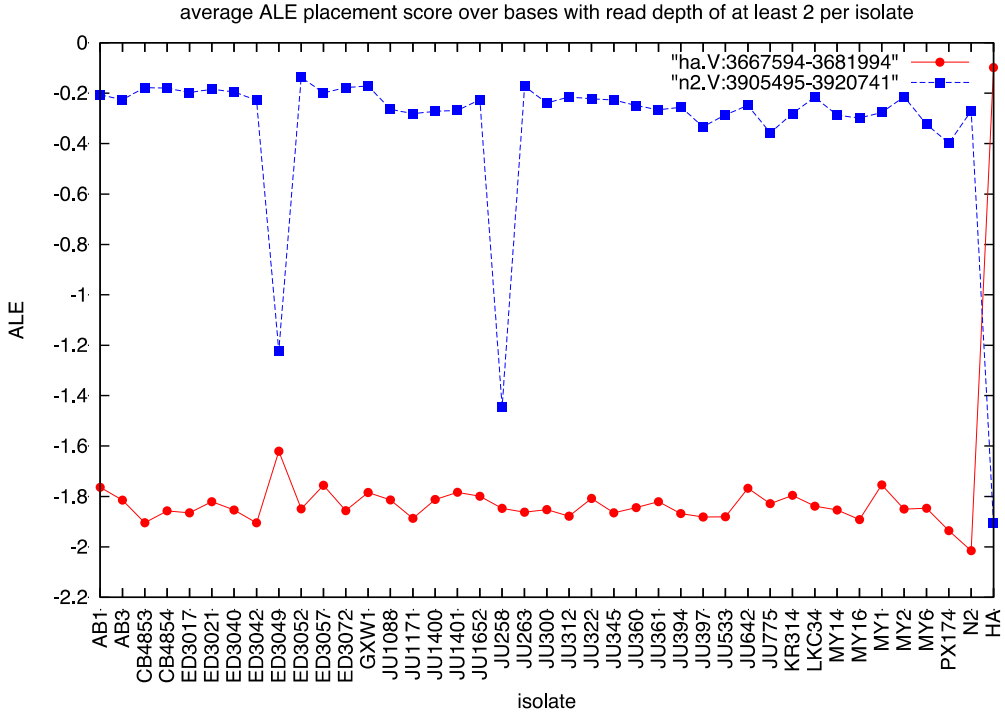

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:2445201-2455014"
"h2.V:2580971-2592469"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

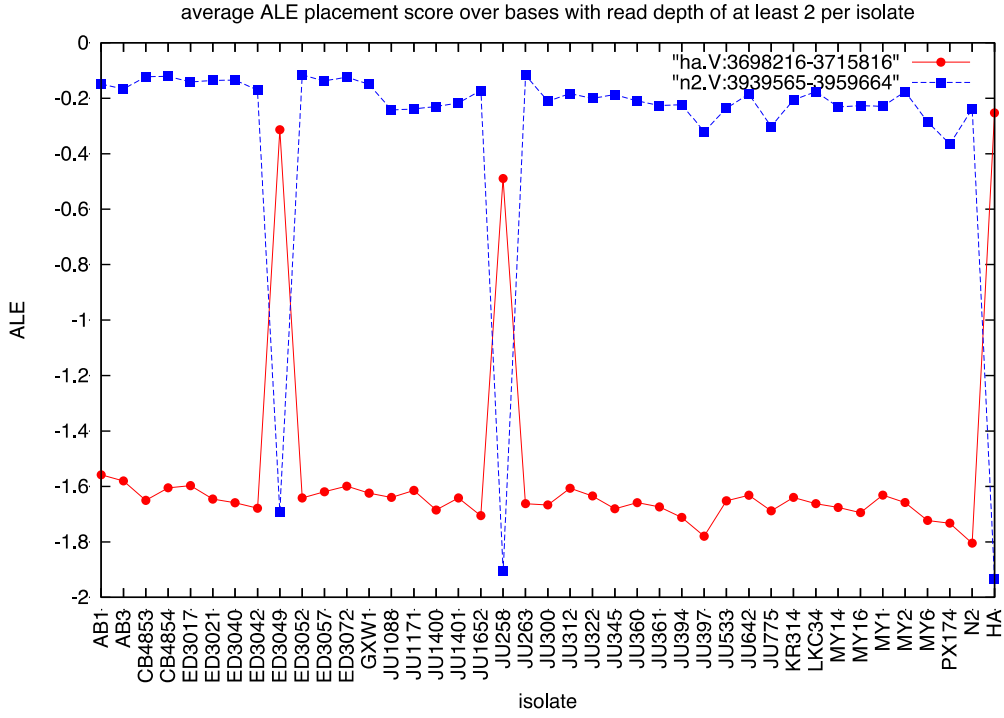

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:2665323-2683253"
"n2.V:2802647-2821695"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

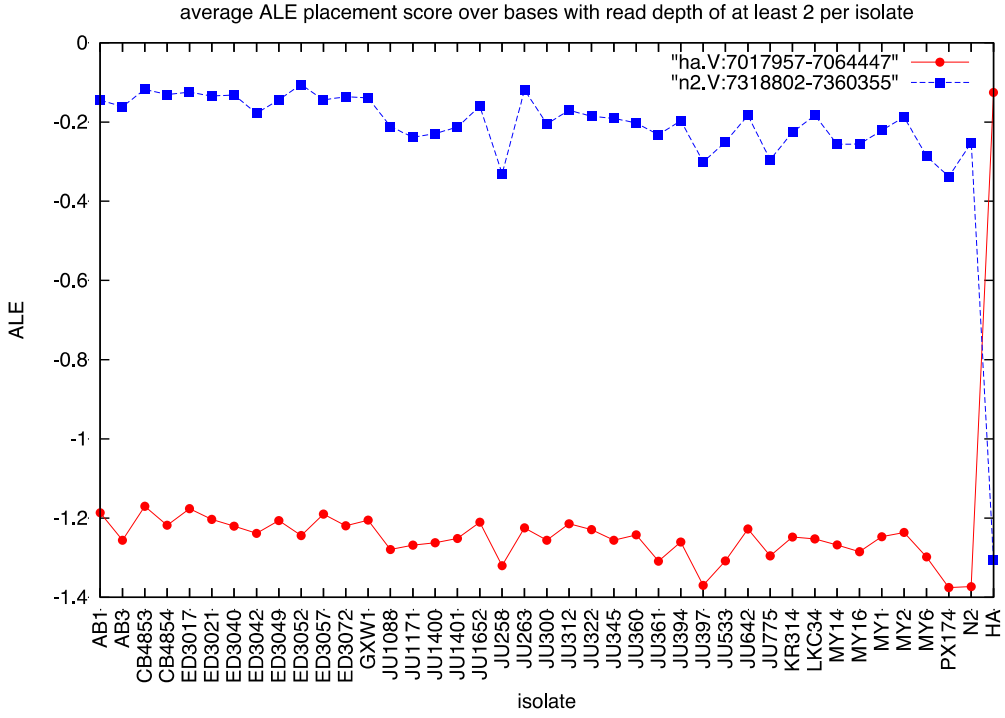

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:3163005-3142615"
"ha.V:3277718-3286975"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
-4.5
-5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

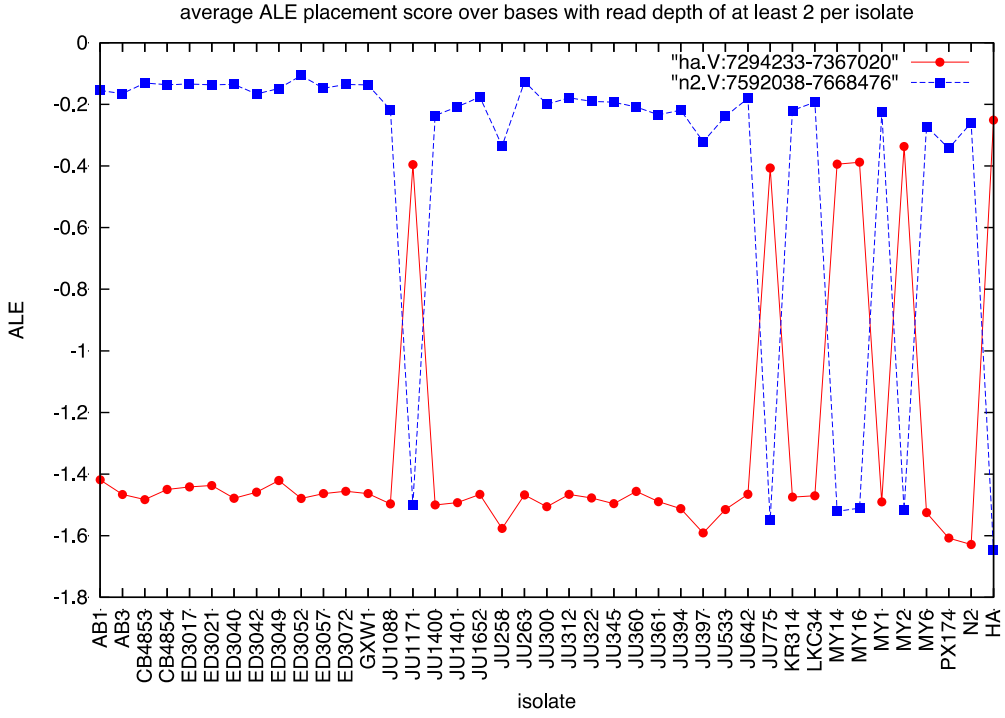

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:3175100-3211048"
"p2.V:3319347-3445836"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

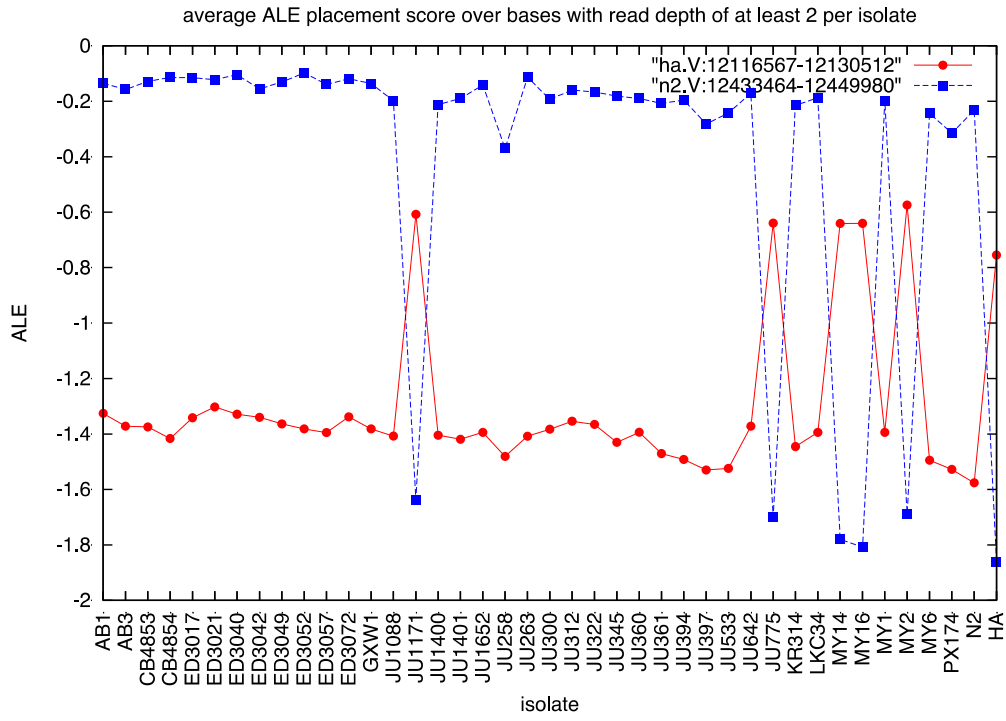

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:3621005-3631859"
"n2.V:3859108-3870050"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

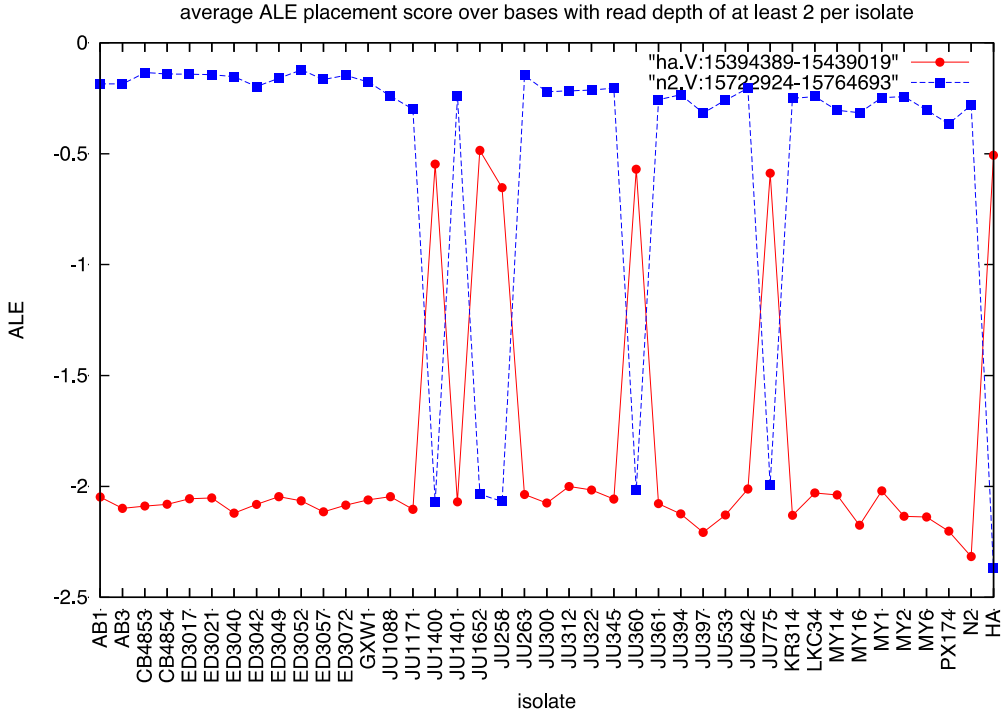

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:3667594-3681994"
"n2.V:3905495-3920741"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
-2.2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

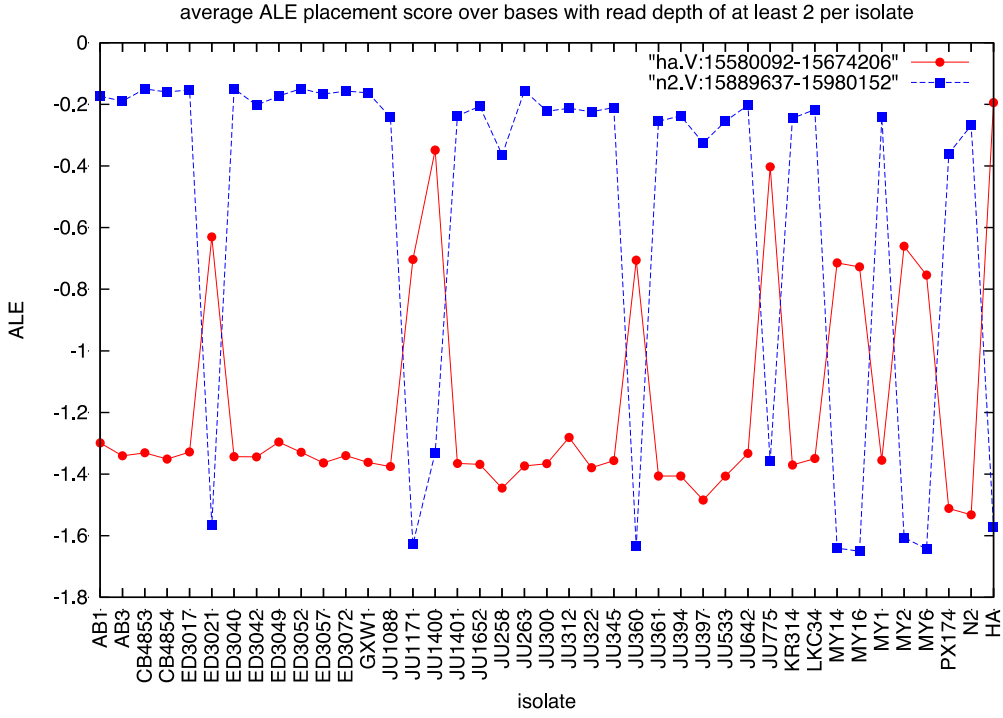

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:3698216-3715816"
"n2.V:3939565-3959664"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

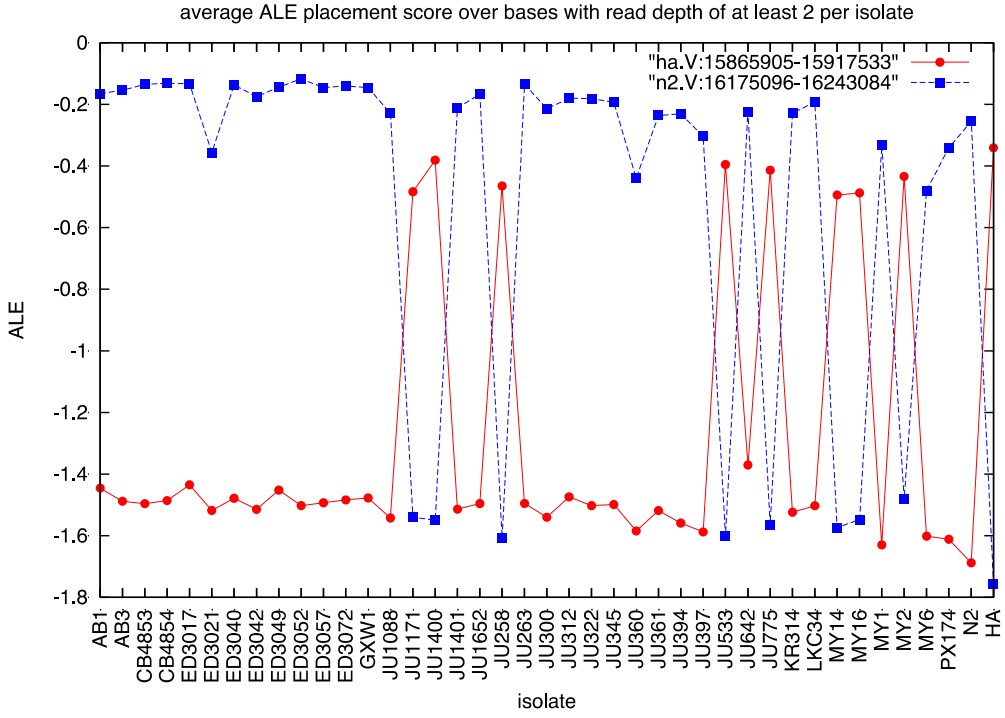

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:7017957-7064447"
"n2.V:7318802-7360355"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

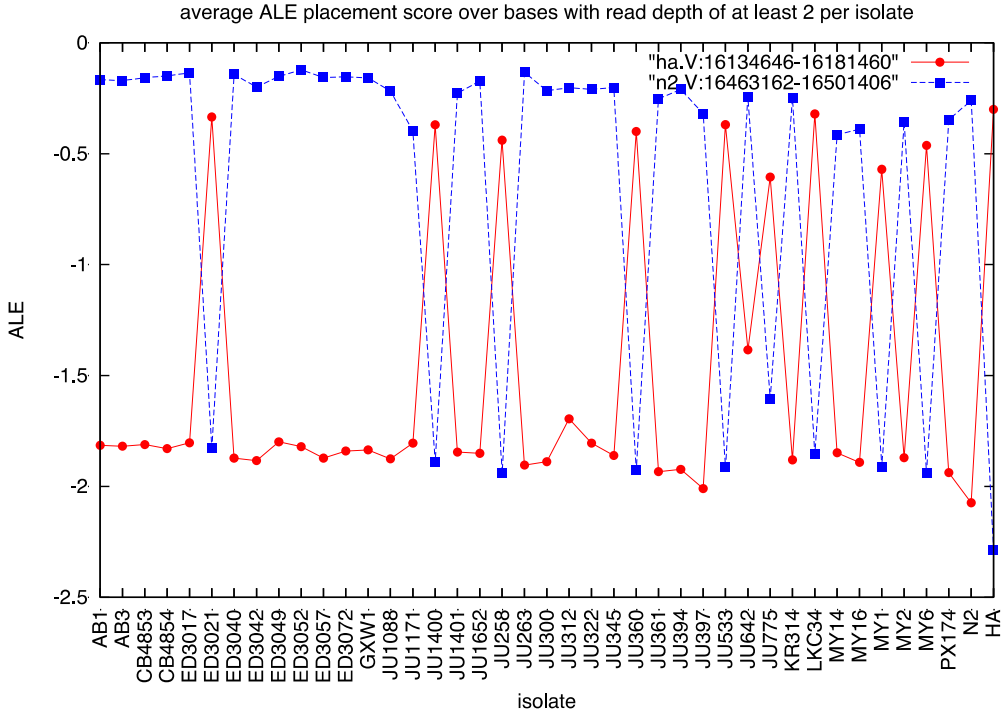

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:7294233-7367020"
"n2.V:7592038-7668476"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

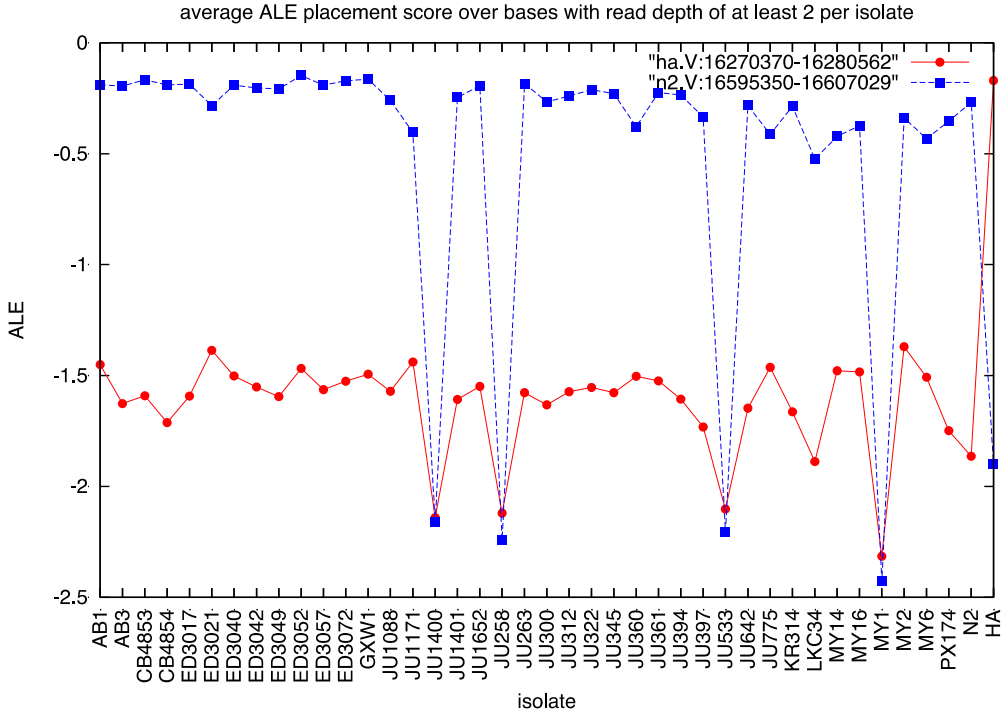

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:12116567-12130512"
"n2.V:12433464-12449980"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
-2
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

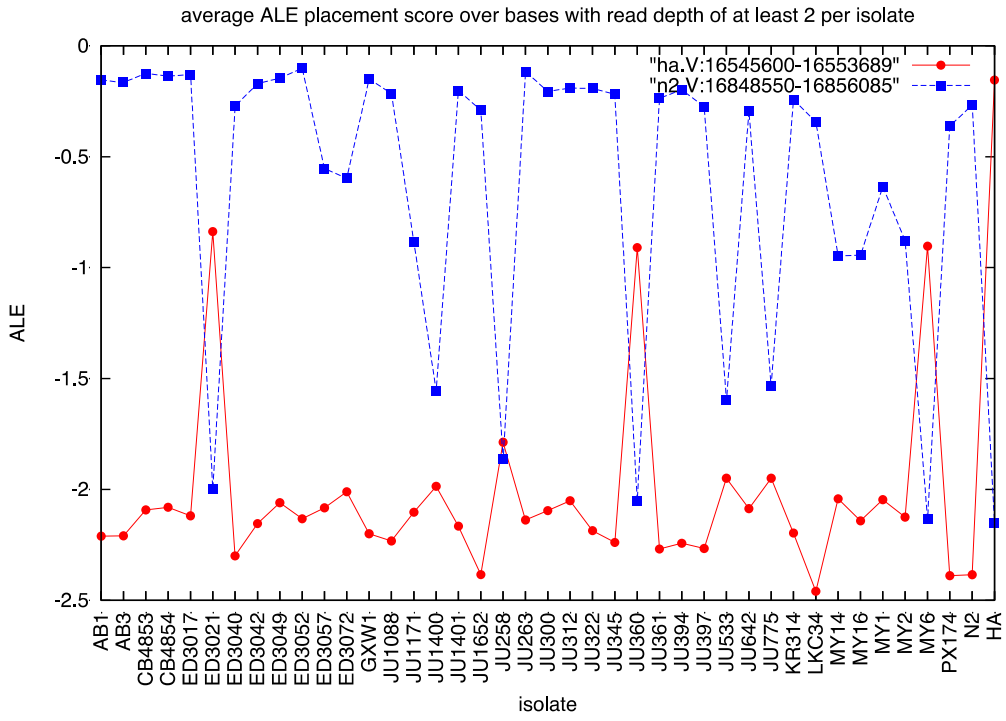

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:15394389-15439019"
"n2.V:15722924-15764693"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

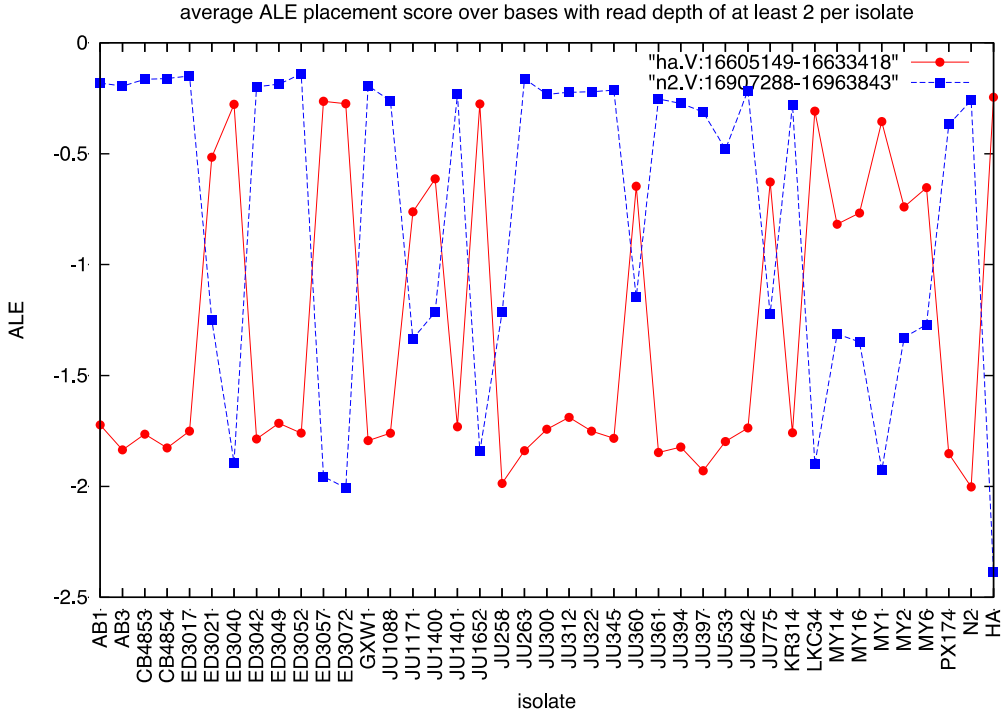

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:15580092-15674206"
"n2.V:15889637-15980152"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

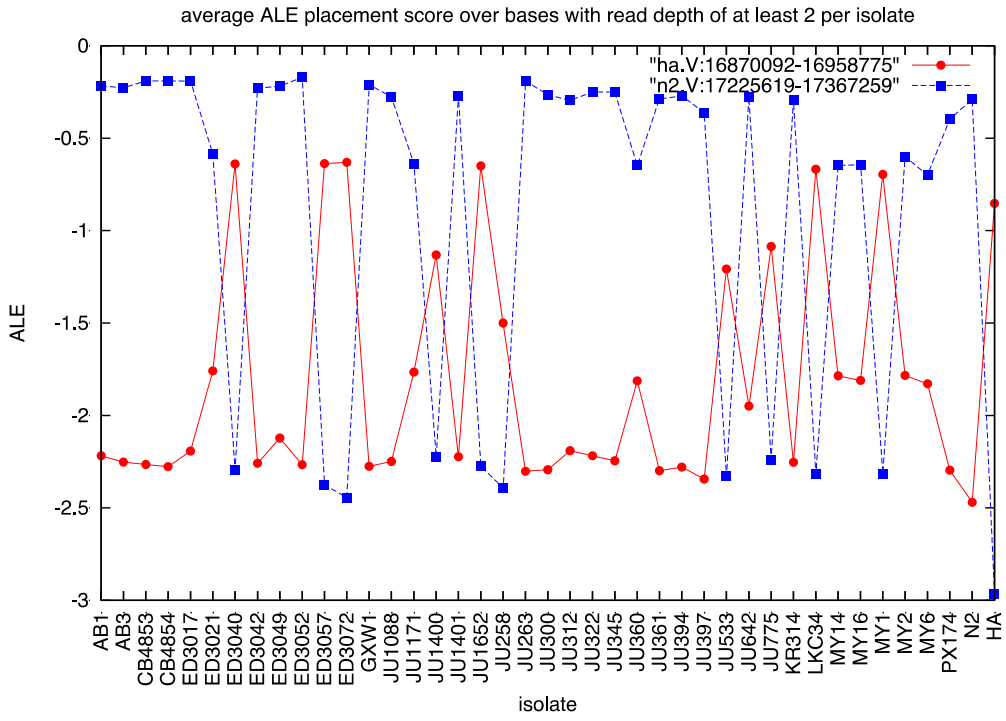

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:15865905-15917533"
"n2.V:16175096-16243084"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

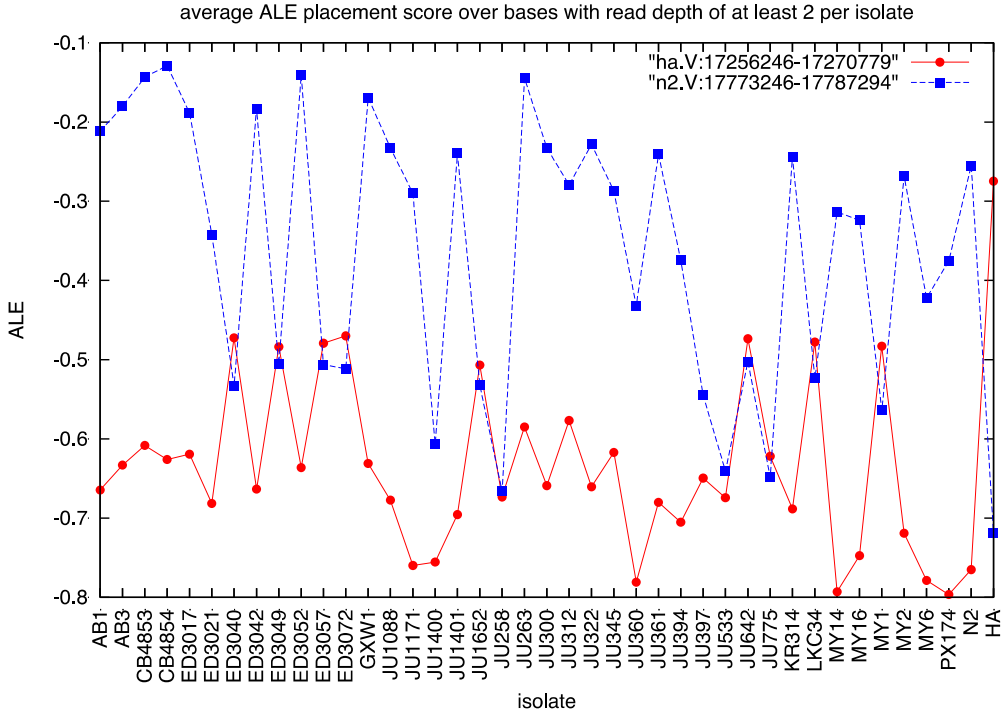

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:16134646-16181460"
"n2.V:16463162-16501406"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

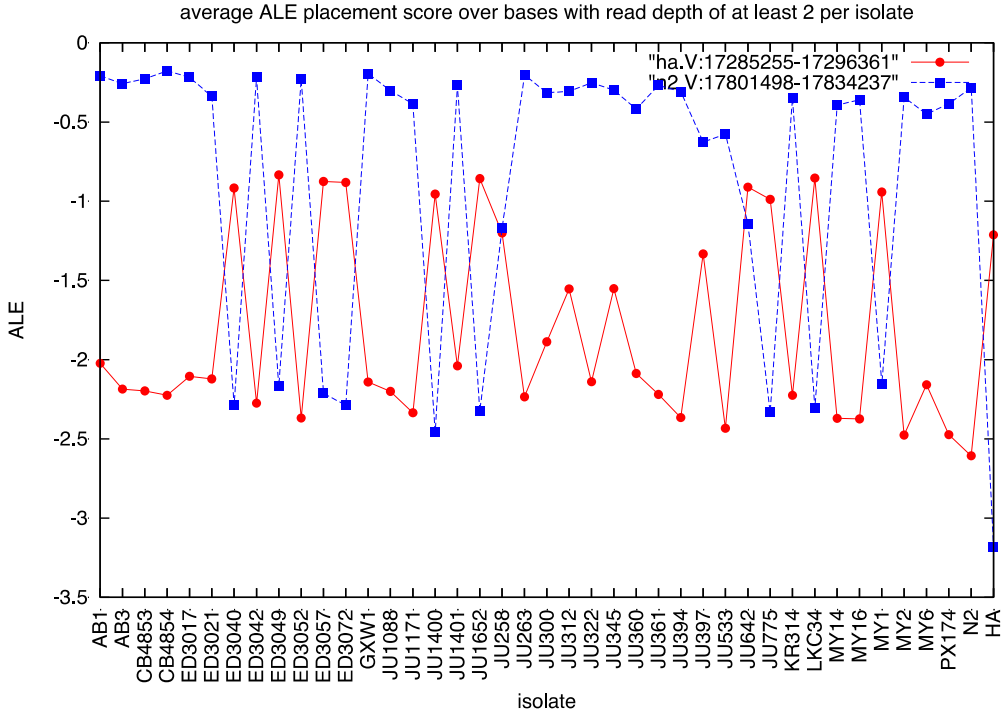

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:16270370-16280562"
"n2.V:16595350-16607029"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

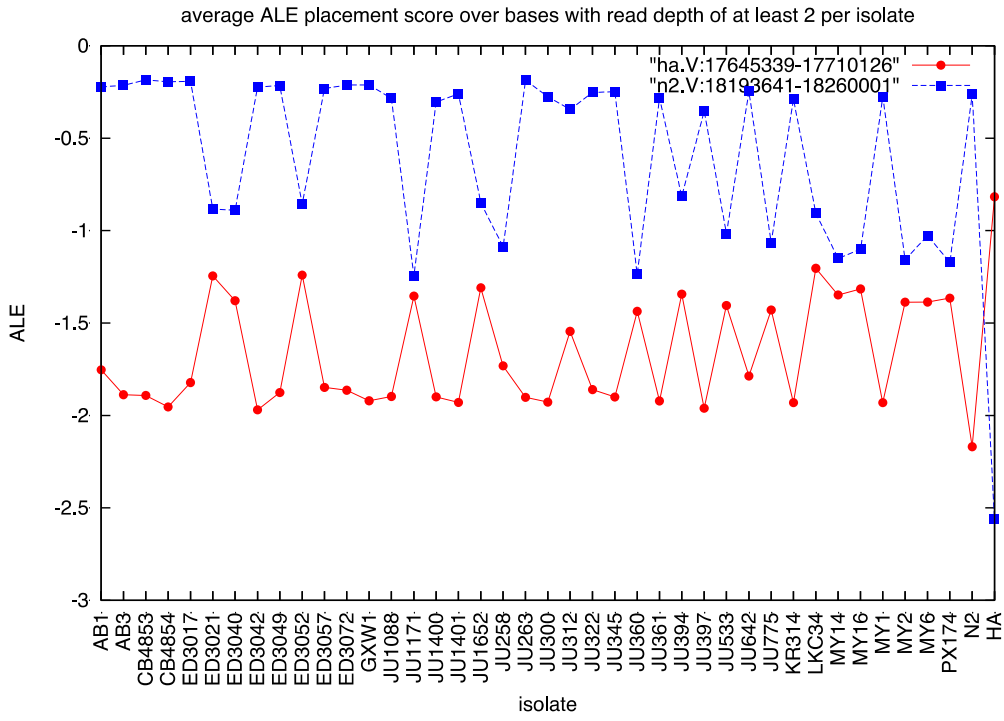

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:16545600-16553689"
"n2.V:16848550-16856085"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

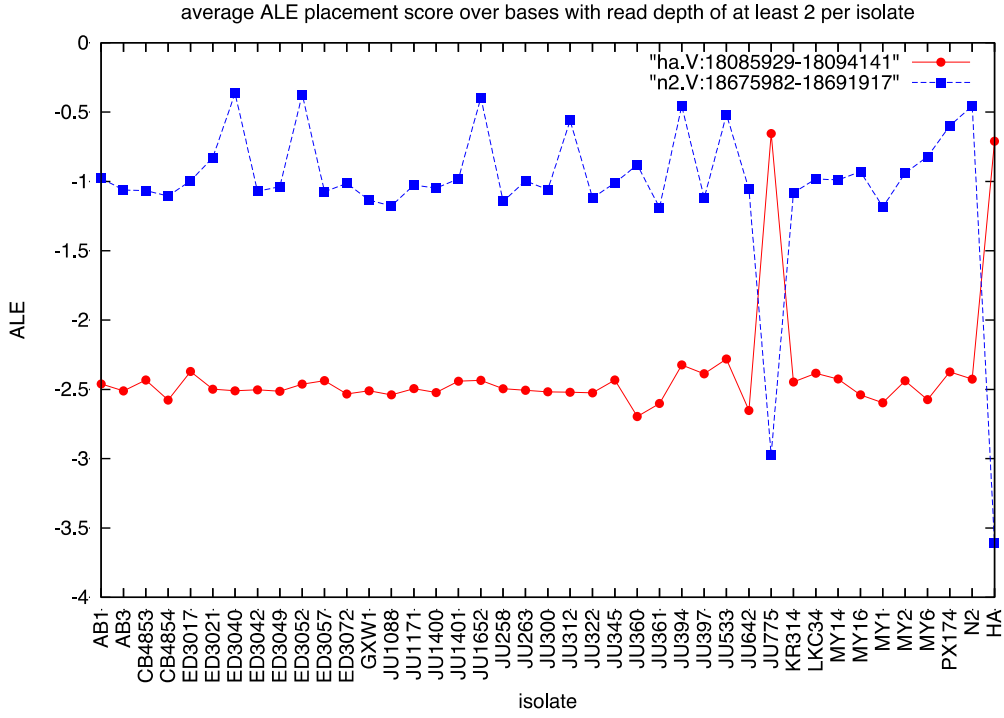

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:16605149-16633418"
"n2.V:16907288-16963843"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

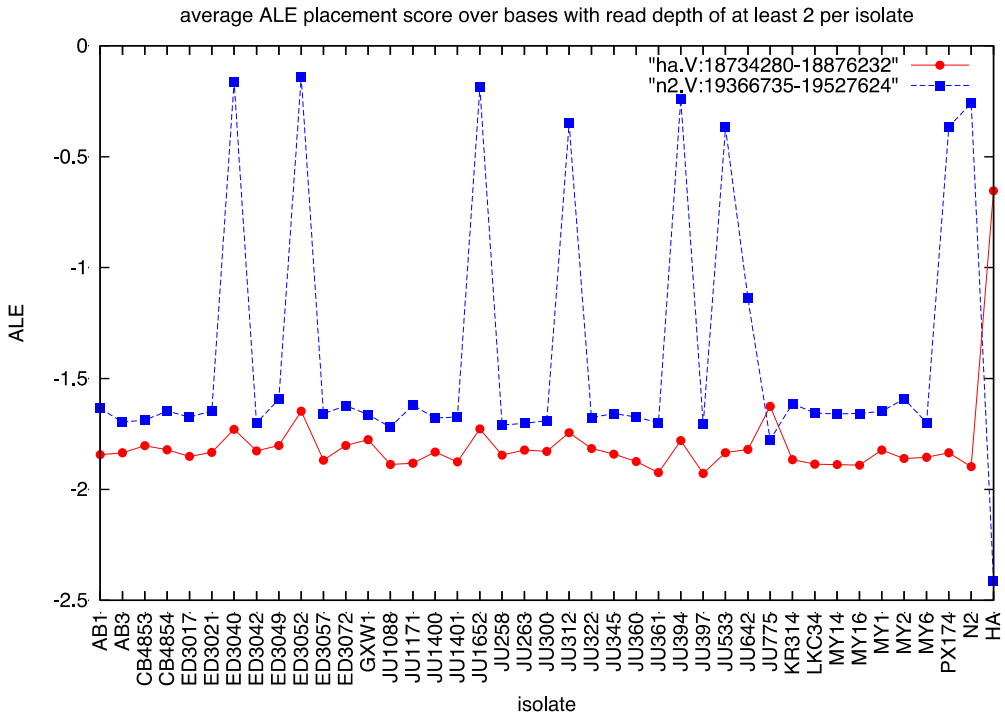

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:16870092-16958775"
"n2.V:17225619-17367259"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

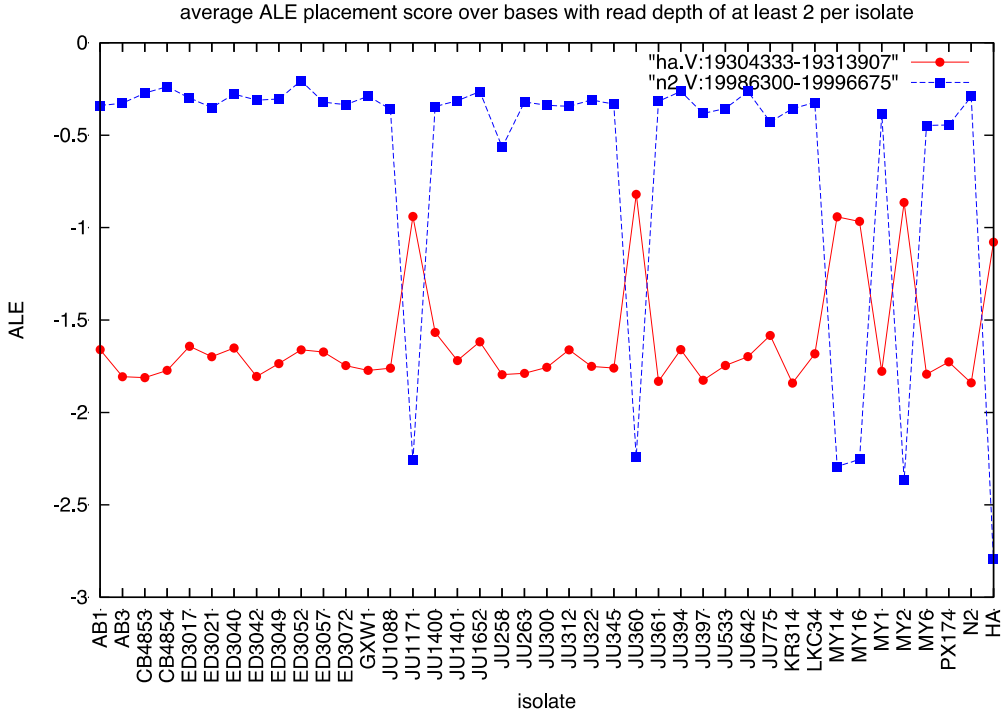

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:17256246-17270779"
"n2.V:17773246-17787294"
ALE
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

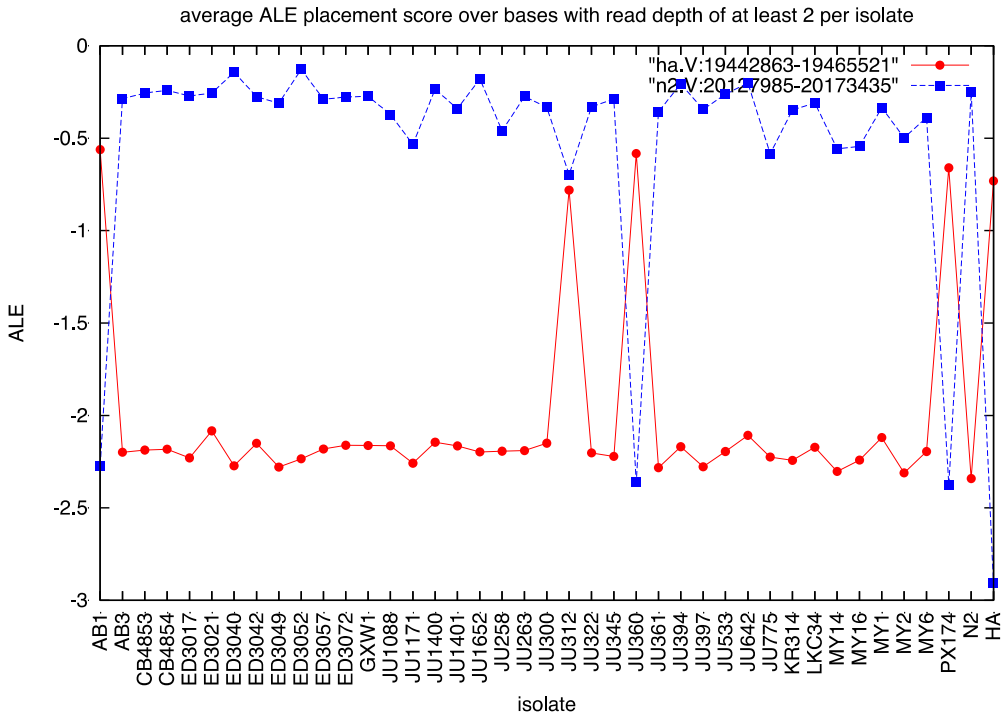

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:17285255-17296361"
"ha.V:17801498-17834237"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

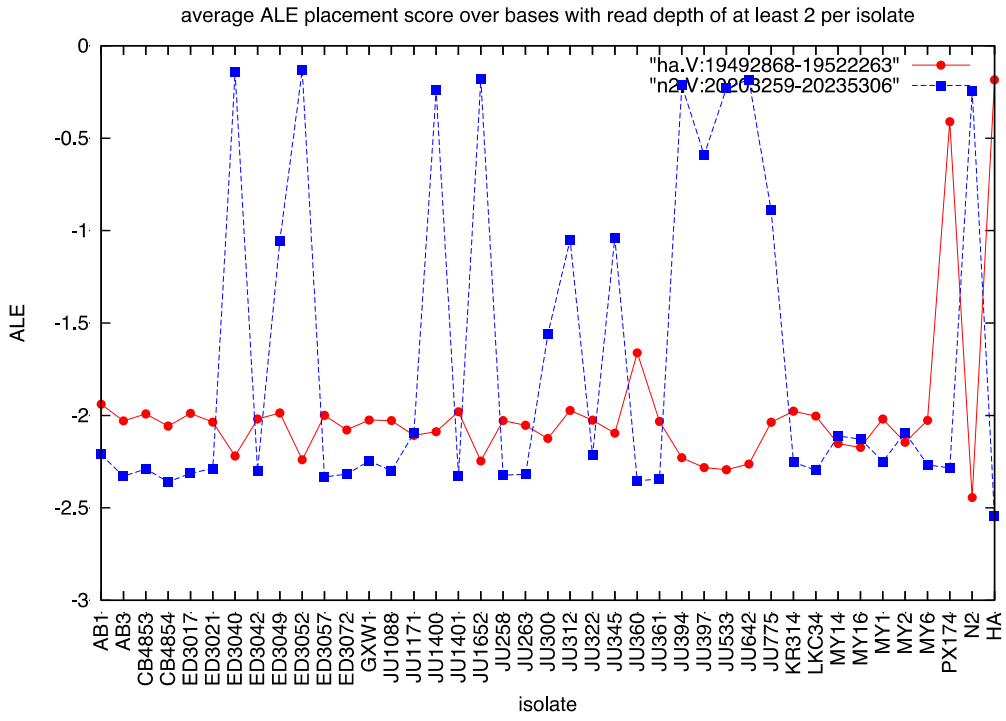

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:17645339-17710126"
"n2.V:18193641-18260001"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

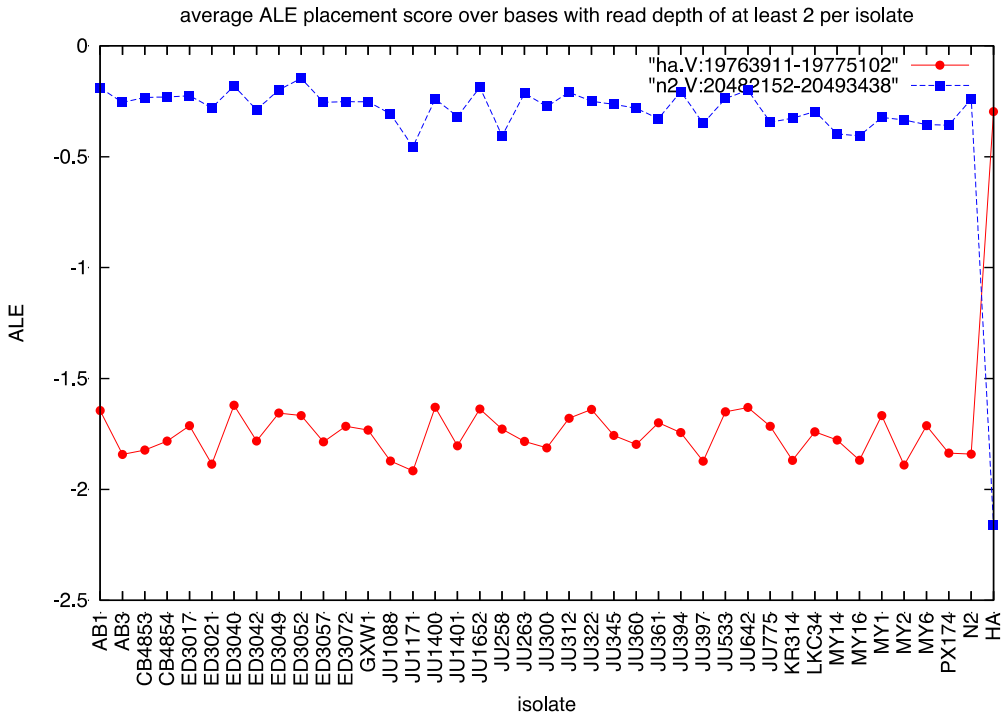

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:18085929-18094141"
"n2.V:18675982-18691917"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

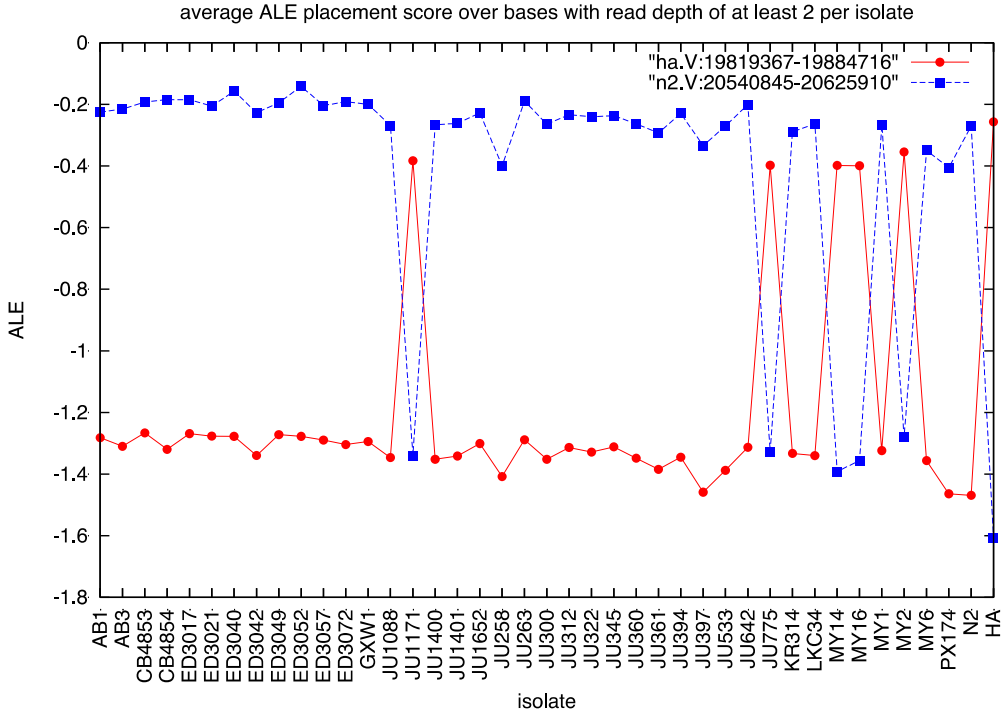

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:18734280-18876232"
"n2.V:19366735-19527624"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

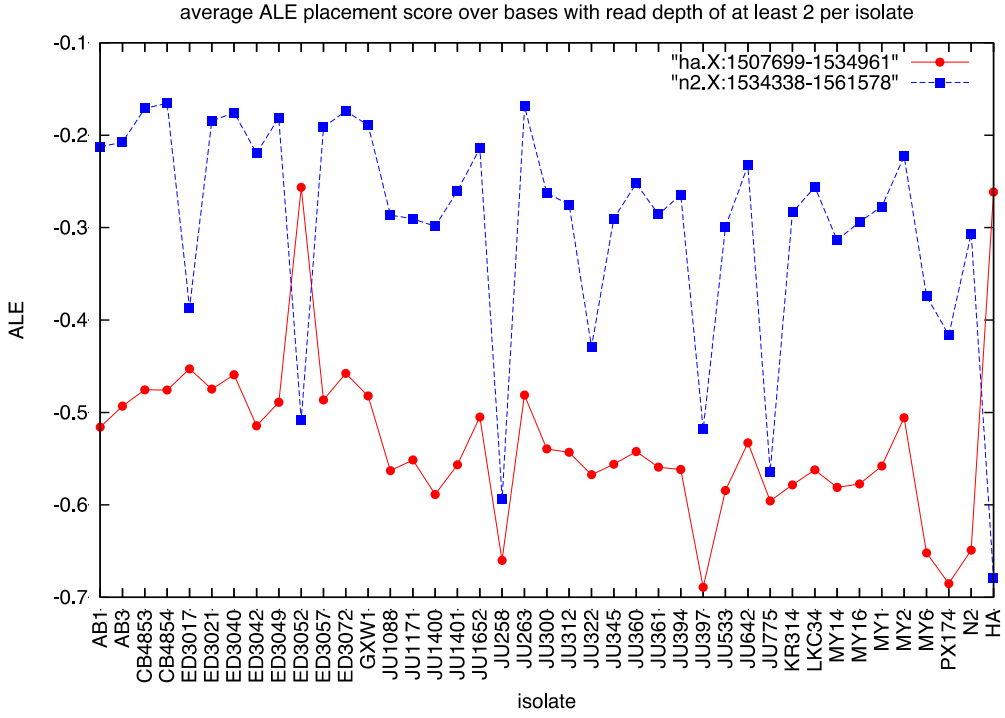

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:19304333-19313907"
"n2.V:19986300-19996675"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

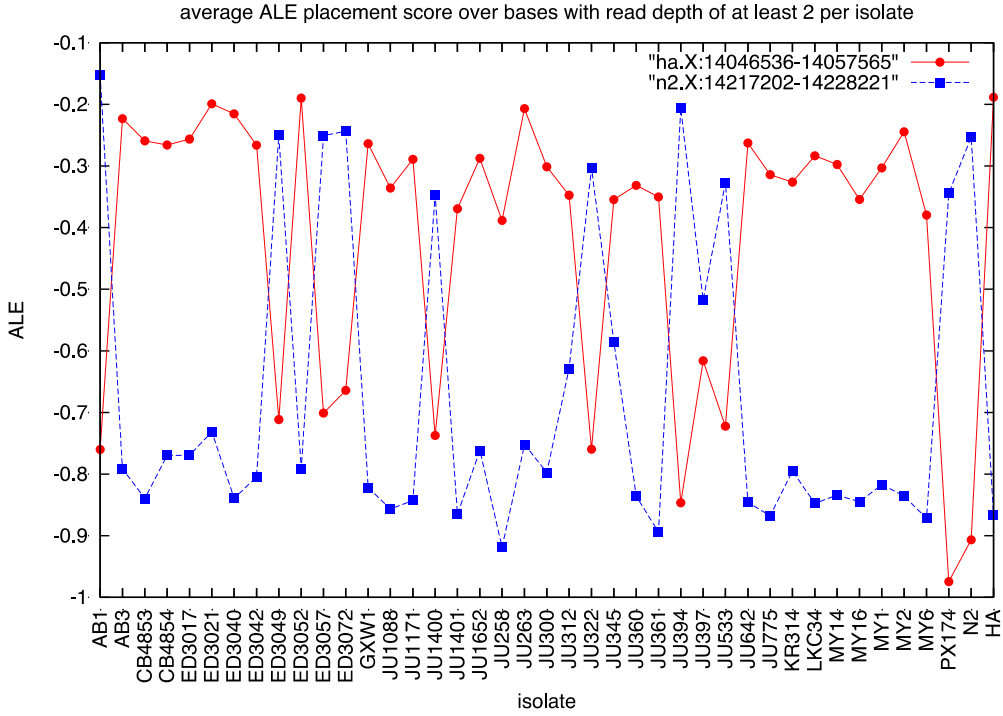

average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:19442863-19465521"
"n2.V:20127985-20173435"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate


average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:19492868-19522263"
"n2.V:20203259-20235306"
ALE
0
-0.5
-1
-1.5
-2
-2.5
-3
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate


average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:19763911-19775102"
"n2.V:20482152-20493438"
ALE
0
-0.5
-1
-1.5
-2
-2.5
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate


average ALE placement score over bases with read depth of at least 2 per isolate
"ha.V:19819367-19884716"
"n2.V:20540845-20625910"
ALE
0
-0.2
-0.4
-0.6
-0.8
-1
-1.2
-1.4
-1.6
-1.8
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate


average ALE placement score over bases with read depth of at least 2 per isolate
"ha.X:1507699-1534961"
"n2.X:1534338-1561578"
ALE
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate


average ALE placement score over bases with read depth of at least 2 per isolate
"ha.X:14046536-14057565"
"n2.X:14217202-14228221"
ALE
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1
AB1
AB3
CB4853
CB4854
ED3017
ED3021
ED3040
ED3042
ED3049
ED3052
ED3057
ED3072
GXW1
JU1088
JU1171
JU1400
JU1401
JU1652
JU258
JU263
JU300
JU312
JU322
JU345
JU360
JU361
JU394
JU397
JU533
JU642
JU775
KR314
LKC34
MY14
MY16
MY1
MY2
MY6
PX174
N2
HA
isolate

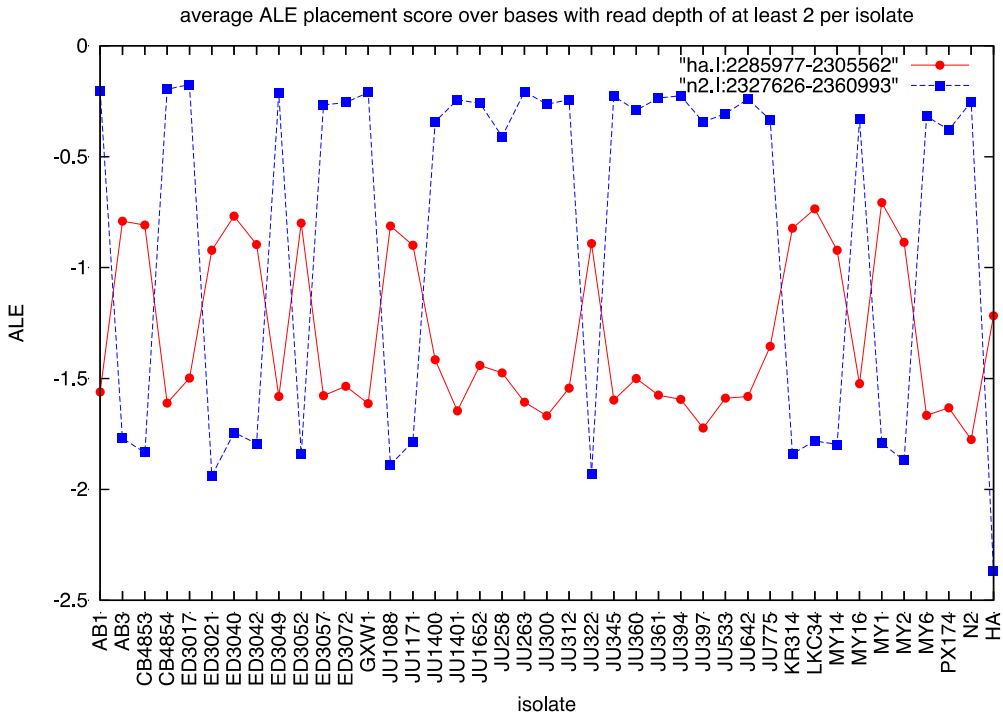


Figure S10 Match of sequence reads from 39 different wild isolates, N2 and CB4856 (HA) against the N2 genome and the CB4856 genome in all 61 divergent regions. The extent of the match is measured by ALE placement scores averaged across the region, with larger negative scores indicating a poorer fit. See Figure S4 for more details of interpretation.

File S1

LASTZ alignment between *C. elegans* N2 (WS230) and CB4856

File S1 is available for download as a text file at http://dx.doi.org/10.5061/dryad.1k8kq

File S2

SNVs identified in *C. elegans* N2 (WS230) and CB4856 comparisons with corresponding sequence features

File S2 is available for download as a text file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1

File S3

INDELs identified in *C. elegans* N2 (WS230) and CB4856 comparisons with corresponding sequence features

File S3 is available for download as a text file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1

File S4

Summary counts of SNV and INDELs identified in *C. elegans* N2 (WS230) and CB4856 comparisons

File S4 is available for download as an excel spreadsheet file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1

File S5

Effects of variation on protein coding genes from *C. elegans* N2 (WS230) and CB4856 comparisons

File S5 is available for download as a text file at
www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175950/-/DC1

File S6

Materials and Methods

Creating a Hawaiian reference

In this section we expand upon the overview presented in the main text.

Iterative editing using variants: As a first step in generating a CB4856 reference sequence, we used the *C. elegans* genome assembly N2 (WS230) as a template, and iteratively edited it based on the variants detected with aligned CB4856 reads from the Princeton data set. To identify SNVs, indels (gapped insertions/deletions) and other variants detected with "split" reads (non-contiguous read alignments suggesting complex and/or intermediately sized variants), we modified our previously published pipeline (THOMPSON *et al.* 2013). We also inserted new sequence based on the consensus behavior of clipped reads. The SNVs, indels, split-read variants and clipped read extensions were integrated into a single pipeline as described below.

Step 1: In an initial step, we corrected the alignment of reads that were falsely clipped (Figure S1A) or improperly aligned to the reference (Figure S1B) that matched other reads at the same position. In the first instance, reads adjacent to an alignment gap had been clipped by the aligner because the unmatched portion of the read was too short to be confidently split. However, in the presence of split reads with longer extensions across the gap, and where the clipped portion matched that of other split reads with longer extensions, we altered the alignment of those reads by modifying the CIGAR code (the description of the alignment of the read to the reference sequence in BAM alignment files as described in the samtools package (LI *et al.* 2009); reads 1 and 5 in Figure S1A). In the second instance, reads adjacent to an alignment gap might have a short, often SNV-enriched match against the immediately adjacent sequence. Again in the presence of other split reads with longer overlaps, and where the extended alignment matched the other gapped or split reads, we altered the alignment of those reads by modifying the CIGAR code (reads 1 and 5 in Figure S1B).

Step 2: With spuriously clipped reads corrected, at all remaining sites containing reads clipped on either side, we derived a consensus sequence for the clipped sequence and stored it for possible integration into the reference.

Step 3: For each base, we called a SNV if the coverage was at least 4x (the equivalent of 3.8 standard deviations below the mean in the CB4856 Princeton dataset), and if the phred-weighted coverage of the SNV allele (defined as the sum of the phred-scaled base quality scores supporting the allele, divided by the sum of the phred-scaled base quality scores of all aligned bases at the site) was >= 0.8. We then altered the base at that position in the reference.

Step 4: We called short indels detected within the alignment of single, unsplit reads if the coverage was at least 4x and the phred-weighted coverage of the indel allele was >= 0.8. Again, we altered the reference sequence accordingly.

Step 5: We called larger deletions at sites of split reads where the donor/acceptor pairs in the split-read relationship were reciprocally detected (i.e. left and right clipped sites predicted the same variant). We required >=5x coverage and the fraction of reads being right-clipped and left-clipped to be >=0.8. We introduced either a deletion, a deletion with an associated insertion, or a small duplication/insertion less than the size of a read length, as previously described (THOMPSON *et al.* 2013). We altered the reference sequence accordingly. Large duplications predicted by the split-read pipeline were not introduced, nor were predicted inversions or translocations.

Step 6: For both left- and right-clipping sites, where a reciprocally defined split-read deletion could not be called and thus were presumptive insertion sites, we integrated the consensus of the stored clipped sequence from step 2 above as an insertion as shown in Figure S1. This allowed us to introduce novel sequence into the CB4856 reference, anchored to existing sequence derived from the N2 genome. After 20 total cycles of reference editing this approach successfully resolved many CB4856 insertions with respect to the N2 reference, particularly where repetitive sequence was not involved.

After 20 cycles of iteratively editing the WS230 genome with CB4856 alignments, further cycles resulted in few changes. The number of variants of each type introduced by cycle can be seen in Table S1.

Contig Generation: To make further improvements to the CB4856 reference, we generated a *de novo* short-read assembly of the CB4856 Princeton data set with JR-Assembler (CHU *et al.* 2013). We set the insert size parameter for JR-Assembler at 321, the mean insert size of the library when aligned against the N2 reference. After several tests of the software we set the other parameters: (1) to include low-complexity regions in the assembly, (2) not to trim 5' bases from the reads, (3) to use the entire read as the assembly read length, (4) to enable base correction, (5) to connect overlapping paired-end reads, and (6) to use a resampling read shift of 5bp. We used default kernel parameters (minimum overlap length = 30bp, maximum overlap length = 45 bp, minimum remapping ratio = 0.2, minimum contig length = 300 bp) and both constructed a contig order and filled small gaps.

These parameters gave 12,155 contigs with an N50 of 18,813 bp. To remove possible false joins in these contigs, they were then processed with REAPR to split contigs at locations where alignments generated with smalt (sanger.ac.uk/resources/software/smalt/, the optimal aligner the ALE pipeline described below) generated an automatic fraction coverage distribution error (see HUNT *et al.* 2013 for details). REAPR introduced 1,880 contig breaks in 1,539 contigs, yielding a set of 14,167 contigs with an N50 of 15,785 bp.

Contig Integration: *Initial Alignment, RIL/IL Region Delineation and Usage.* We aligned the JR contigs to the cycle-20 reference sequence using cross_match (a Smith-Waterman alignment tool, P. Green, unpublished) on default settings, but several contigs showed ambiguous locations, particularly in regions that initially had poor read coverage. To derive an independent assessment of a contigs location in the genome, we exploited WGS data sets obtained with the SOLiD platform from 49 RILs and 60 ILs (Table 1; (LI *et al.* 2006; DOROSZUK *et al.* 2009); SRA PRJEB7445) as well as a subset of the CB4856 data sets (WICKS *et al.* 2001; GRISHKEVICH *et al.* 2012; THOMPSON *et al.* 2013). We reasoned that good coverage of the JR contigs with reads from the RILs and ILs would require the presence of a CB4856 segment corresponding to that contig in the IL or RIL. In turn the CB4856 variants in the ILs and RILs would define the CB4856 segments contained in those lines. Finally, the segment shared between all strains yielding good coverage of the JR contig would localize the contig in the genome. To determine which regions in each IL or RIL originated in CB4856 and which originated in N2, we aligned the reads to the WS230 reference in colorspace with bwa-0.5.9, then merged the read alignments with samtools 0.1.18 (for coverage statistics see Table S2) and determined the SNV content of each strain across the genome compared to a list of 183,509 SNVs called independently in all three of the Washington, Technion, and Princeton CB4856 datasets. For each RIL/IL, we examined the fraction of SNV locations that contained only CB4856 alleles, moving across each chromosome in 1kb increments (Figure S2). We excluded sites at which only the CB4856 allele was present in more than 30 strains, likely reflecting an error in the WS230 reference or problems with alignment. RIL/IL regions were annotated as originating from CB4856 when more than 150 consecutive 10 kb bins in 1 kb steps (150 kb) contained nonzero fractions of exclusively CB4856 SNV alleles, at least one bin in the region contained 100% CB4856 alleles, and at least 50 successive 10 kb bins in 1 kb steps (50 kb) contained no CB4856 alleles to either side (or were telomeric). On average each IL contained 1.6 CB4856 regions per line, and each RIL contained 9.0 regions per line.

In parallel, we aligned the SOLiD reads from each line against the JR contigs in colorspace, again using bwa-0.5.9 on default settings and merging with samtools. For each contig, two distributions were calculated across the set of 60 ILs: (a) the coverage of each IL when aligned to the contig, normalized to the total coverage in the IL as defined in Table S2, and (b) the fraction of the contig covered to a depth of >= 1x. Based on the mean and standard deviation for each of these statistics, each contig was given two Z scores. We heuristically determined that if either Z score was greater than 3 for a given IL, or if one Z score was greater than 2 and the other greater than 1, the contig was likely to trace its origin to the defined CB4856 regions in that strain. For example, contig1397-1 (aka, contig1397_start3075_9876.Reapr1) was covered at a substantially higher fraction of its bases, and to a substantially higher depth, in IL lines ewir064 and ewir066, which shared CB4856 sequence only on chromosome V (Figure S3). The IL-defined location of the contig that was consistent with the

position of the contig aligned to the iteratively edited reference (+/- 50 kb to account for differences in coordinates in the two references) was then accepted at the correct location.

Contig Integration: *Contig Placement***.** 11,254 of the JR contigs (79.4%) aligned fully and contiguously to the iterative CB4856 draft genome with cross_match, an improvement from 10,216 (72.1%) fully aligning to the WS230 genome. Of the remaining 2,913 contigs, 2,716 produced at least one high quality (HQ) alignment (having the maximum cross_match alignment score of 254) to the iteratively edited reference. 1,902 of these produced more than one HQ alignment, often to a given region but sometimes dispersed across chromosomes in the iteratively edited reference, particularly where the iterative extension of sequences failed to yield contiguous coverage, or where the contig contained repeated sequences. 123 contigs produced more than five initial HQ alignments. To improve the sequence in these regions that had proved recalcitrant to modification through our iterative editing procedure, we sought to replace the problematic areas with the *de novo* assembled sequence. The complicated relationships of contig alignments in some of these regions, many of which were flagged by REAPR and ALE to indicate remaining assembly problems, led us to develop a series of procedures to resolve discrepancies where we could, and to use the JR contig information conservatively.

For those contigs with HQ alignments across different chromosomes, we used the IL mapping procedure described above to determine which alignments were consistent with the placement of the contig in the ILs. Of 311 such contigs, we were able to define a genomic region for 177: in 44 the region predicted by the IL analysis did not contain any of the candidate alignments, and in another 90 the ILs failed to identify a genomic region, primarily because the CB4856 sequences in these regions were too similar to N2 to cause a drop in mapping density. After resolving chromosomal locations in this way, 260 contigs yielded HQ alignments to different strands (38 of these had been localized with the IL analysis), perhaps the result of local inversions between N2 and CB4856. To determine the orientation of these contigs for integration, we used a simple heuristic: if strand A accounted for >= 80% of the contig, or if strand A generated more total alignments and accounted for >= 60% of the contig, we selected strand A. For 93 contigs, the strand was ambiguous by the above criteria and these contigs were left unresolved.

For those 2,489 contigs with HQ alignments that could be assigned to a region and a strand, we added any additional alignments (i.e. even those with lower mapping quality) to our set of HQ alignments provided they shared the same chromosome and strand. We then defined regions in which the alignments to the reference were in the same order as in the contig and were not interrupted by the unique alignment of another contig segment. If the whole contig aligned to a single chromosomal span and contained five or more alignments, we replaced the whole region with the JR contig (221 contigs totaling 2,384,154 bp, replacing 2,538,749 bp of the edited sequence). For other contigs with fewer discrepancies, we simply

replaced the segments surrounding any gaps with the corresponding segment of the JR contig, introducing insertions, deletions or duplications present in the JR contig (1,543 replacements from 1,264 contigs). 333 segments gave inconsistent results and were not used. Integrating these changes into one more iteration of our variant calling pipeline, we were able to identify 1,574 additional gapped indels, supported by the alignment of the contigs to the reference, that were present in our earlier alignments but had coverage below the thresholds of our variant calling pipeline so had not been replaced in the early rounds of iterative editing.

Residual N2-only Sequence Removal: In addition to integrating REAPR-processed JR-Assembler contigs into our final assembly, we also removed stretches of sequence that we determined to be absent from the CB4856 genome. These regions had persisted in the iteratively edited reference because up to this point we had only deleted sequences where we could determine the end points with basepair resolution through consensus alignments (repeated sequence at the ends of the likely deletions were a major problem here). Thus, there remained stretches of original N2 sequence in our CB4856 reference. To identify and remove these sequences we looked for stretches of 100 bp or more that generated zero coverage with smalt. Smalt places multiply mapped reads randomly, ensuring coverage of repeated sequence, which might otherwise be flagged as having no coverage. We then expanded the region both upstream and downstream until coverage rose above two standard deviations less than the mean coverage for the strain. When, not infrequently, such a region ended inside an annotated RepeatMasker (SMIT *et al.* 1996-2010) motif, we extended the region to the end of the repeat. 1,098 regions totaling 1,565,559 bp were removed from the final assembly using the above criteria.

Supplementary Tables

Table S1 Variant calls by iteration

Iteration	SNVs	Indels	Split-Read Variants	Clipped-Read Extensions
1	219787	46674	4905	6856
2	17790	5504	1226	3889
3	8412	1927	543	2461
4	4890	983	280	1618
5	3023	598	130	1130
6	2104	366	77	775
7	1230	206	47	594
8	940	157	27	439
9	734	99	13	341
10	490	66	6	294
11	390	67	5	236
12	321	53	4	185
13	230	33	5	161
14	285	36	4	126
15	213	24	1	102
16	131	18	4	94
17	107	20	1	82
18	111	18	1	71
19	93	14	0	59
20	65	11	1	50

Table S2 CB4856 IL and RIL coverage and region list

Strain (alias)	Type	SOLiD N2 coverage	CB4856 Regions	Region List
ewir001	IL	3.23	3	I:1-1894000 II:13393001-13738000 IV:776001-1489000
ewir002	IL	6.53	1	I:2777001-7909000
ewir004 (ewir073)	IL	9.87	1	V:14933001-16302000
ewir005	IL	7.41	1	I:6280001-10873000
ewir007	IL	7.64	1	I:3401001-10873000
ewir009	IL	7.68	1	I:9135001-11315000
ewir010 (ewir039)	IL	7.08	1	III:6444001-11217000
ewir011	IL	6.42	3	I:9135001-14682000 II:13195001-13400000 V:16546001-20913000
ewir012	IL	7.13	1	I:9135001-15059000
ewir016	IL	4.61	1	I:11379001-15058000
ewir017	IL	5.52	1	I:12433001-15059000
ewir019	IL	3.61	2	II:1-2224000 V:18774001-19010000
ewir020	IL	5.70	2	II:1-2915000 V:18774001-19010000
ewir021	IL	11.00	1	II:1-3598000
ewir023	IL	4.17	1	II:2850001-11030000
ewir024	IL	11.35	0	
ewir025	IL	5.21	1	II:12610001-13081000
ewir026	IL	6.74	1	II:11794001-13907000
ewir027	IL	6.41	1	II:12262001-15268000
ewir028	IL	6.27	1	III:1-2461000
ewir029	IL	8.75	1	III:616001-2007000
ewir030	IL	6.88	1	III:1-3096000
ewir032	IL	5.77	1	III:1857001-2948000
ewir034	IL	5.79	2	III:2833001-5473000 III:11580001-12156000
ewir035	IL	10.01	1	III:2571001-9675000
ewir039	IL	7.37	1	IV:9068001-13348000
ewir040 (ewir041)	IL	7.09	1	III:8034001-11308000
ewir042	IL	7.19	1	III:9015001-11436000
ewir043	IL	6.54	1	III:12141001-13772000
ewir044	IL	7.78	1	III:11580001-13772000
ewir045	IL	7.25	2	IV:168001-308000 IV:388001-2272000
ewir046	IL	8.62	1	IV:524001-2740000
ewir048	IL	3.78	1	IV:1164001-3485000
ewir050	IL	19.08	1	IV:2272001-4403000
ewir051	IL	7.15	2	IV:1418001-4403000 IV:4472001-6458000

ewir052	IL	11.05	1	IV:1418001-13347000
ewir054	IL	8.94	1	IV:9068001-13349000
ewir055	IL	7.68	1	IV:9845001-13348000
ewir058	IL	4.77	3	IV:10259001-10442000 IV:10507001-10767000 IV:10971001-16248000
ewir060	IL	5.80	1	IV:12683001-16599000
ewir063 (wn291)	IL	5.46	5	II:1-12266000 IV:4472001-16285000 IV:16346001-17482000 X:2572001-6374000 X:15009001-15160000
ewir064 (wn292)	RIL	4.94	5	II:13395001-15268000 IV:14652001-17482000 V:1-8766000 V:8927001-13377000 X:1-1816000
ewir066	IL	8.05	2	V:607001-2454000 V:2557001-5152000
ewir067	IL	7.29	1	V:3857001-6578000
ewir068	IL	9.07	1	V:3857001-13403000
ewir070	IL	5.17	1	V:10577001-20913000
ewir071	IL	7.76	1	V:11379001-13403000
ewir074	IL	5.72	2	II:13076001-14184000 V:14659001-18407000
ewir075 (wn041)	RIL	9.31	5	II:12178001-13691000 IV:994001-10935000 IV:11017001-13472000 X:3460001-4406000 X:4738001-9210000
ewir077	IL	8.36	1	V:19248001-20913000
ewir078	IL	10.35	1	X:1-1146000
ewir079	IL	6.00	1	X:1-2028000
ewir080	IL	6.51	1	X:1-2750000
ewir081	IL	7.31	1	X:1-9096000
ewir084 (wn293) _original	RIL	3.44	7	I:1-1433000 II:1-11834000 IV:5155001-8027000 IV:8099001-8995000 IV:9076001-10284000 X:5746001-13101000 X:13173001-15160000
ewir084 (wn293) _repeat	RIL	10.25	5	I:1-1435000 II:1-11836000 IV:5155001-10324000 X:5746001-13112000 X:13239001-15159000
ewir086	IL	5.23	1	X:5834001-9094000
ewir087	IL	4.55	1	X:9096001-11038000
ewir088	IL	5.50	3	X:9096001-12038000 X:12117001-13112000 X:13173001-14249000

ewir089	IL	8.05	1	X:13173001-16737000
ewir090	IL	6.15	1	X:13255001-17684000
wn009 (wn294)	RIL	4.33	13	I:1-1558000 I:9204001-9353000 II:1-2783000 II:12217001-15268000 III:2337001-2722000 III:3411001-13772000 IV:777001-8966000 IV:9045001-10082000 IV:10164001-11998000 V:2564001-2798000 V:3120001-8428000 X:2195001-13112000 X:13173001-17705000
wn013	RIL	6.90	7	I:1-1961000 II:1-12005000 III:4300001-11283000 IV:12495001-17482000 V:1-19603000 V:19685001-20913000 X:1730001-17705000
wn020	RIL	7.09	7	I:1-1773000 I:2471001-15059000 II:5674001-13394000 III:1-1014000 III:1556001-1980000 V:1-8821000 V:8904001-20913000
wn021	RIL	8.93	7	I:2830001-15059000 II:1-2550000 III:1-13772000 IV:1-332000 IV:416001-4401000 IV:4474001-15452000 V:15581001-20913000
wn031	RIL	6.67	5	IV:2222001-17482000 V:1-8834000 V:8898001-18070000 V:19157001-20913000 X:1-1658000
wn034	RIL	8.78	6	I:3662001-14838000 III:1-11968000 IV:388001-17482000 V:12009001-17751000 X:10132001-13112000 X:13207001-17705000
wn036	RIL	8.50	0	
wn037 (wn076_repeat)	RIL	4.87	11	I:13868001-15059000 III:1-4745000 III:4812001-6088000 III:6155001-11032000 III:11143001-13772000 IV:388001-686000 IV:776001-12632000 IV:12695001-13335000 V:7192001-8843000 V:8904001-20913000 X:6910001-17701000

wn038	RIL	7.59	12	I:1-1143000 I:3070001-15059000 II:1-2641000 III:1-13772000 IV:68001-556000 IV:639001-2763000 IV:14638001-17482000 V:1-2454000 V:2527001-2957000 V:9010001-20913000 X:1-13112000 X:13173001-17705000
wn041 (ewir076)	RIL	7.36	1	V:18590001-20046000
wn046	RIL	6.61	7	I:6032001-15059000 II:1-15268000 IV:7939001-10935000 IV:10996001-16137000 V:1-2059000 X:1-13112000 X:13173001-15147000
wn049	RIL	5.02	8	I:11783001-15059000 IV:548001-1087000 IV:8841001-8978000 IV:9068001-10769000 IV:10842001-17482000 V:2741001-8851000 V:8921001-20913000 X:15089001-17701000
wn053 (wn054)	RIL	6.08	12	I:11576001-15058000 III:2704001-4158000 III:4220001-6158000 III:6219001-6330000 III:6397001-13767000 IV:1-122000 IV:548001-667000 IV:740001-4053000 V:1-2486000 V:2557001-8834000 V:8904001-10598000 V:14830001-20913000
wn057	RIL	7.22	10	I:1-7909000 II:1-13907000 III:1-2007000 IV:3817001-4389000 IV:4461001-8978000 IV:9041001-11058000 IV:11134001-17482000 X:9101001-13112000 X:13176001-14249000 X:15670001-17684000
wn058	RIL	6.43	7	I:1-1790000 II:2472001-15268000 III:5133001-12697000 IV:1010001-11094000 IV:11155001-17482000 V:3458001-13650000 X:1-10965000

wn068 (wn076)	RIL	5.99	12	I:13857001-15059000 III:1-4154000 III:4239001-6330000 III:6397001-13772000 IV:388001-4401000 IV:4472001-10776000 IV:10842001-11416000 IV:11477001-13342000 V:7192001-8797000 V:8898001-20913000 X:6885001-13112000 X:13176001-17701000
wn070 (wn041)	RIL	8.22	7	II:12005001-15268000 III:11330001-13767000 IV:1-220000 IV:524001-1172000 IV:15083001-17482000 V:14904001-20913000 X:1-8665000
wn071	RIL	6.40	9	I:2771001-15059000 II:1-15268000 IV:2505001-2814000 IV:15517001-17482000 V:3058001-8851000 V:8921001-19140000 X:1-12653000 X:12714001-13112000 X:13173001-13964000
wn072	RIL	5.77	12	I:1-1079000 II:1-5417000 III:5472001-13772000 IV:2530001-4401000 IV:4472001-5450000 IV:5522001-17482000 V:1-8797000 V:8898001-15075000 X:1-538000 X:1183001-3046000 X:6021001-13112000 X:13176001-17693000
wn074 (wn174)	RIL	6.37	11	I:2549001-5126000 I:5191001-15059000 II:13174001-15268000 III:1-6157000 III:6242001-12805000 IV:14933001-17482000 V:1-2454000 V:2557001-8766000 V:8921001-14979000 X:1-13094000 X:13173001-13540000
wn075 (wn067)	RIL	11.84	5	I:1-1901000 III:1-4918000 III:4986001-13772000 V:1-20913000 X:9294001-17701000

wn098	RIL	6.02	12	I:1-2297000 II:1-3302000 II:13040001-14375000 III:8582001-13767000 IV:168001-308000 IV:388001-548000 IV:636001-2420000 V:1-2454000 V:2564001-8843000 V:8921001-20913000 X:1-13112000 X:13173001-15520000
wn105	RIL	6.20	11	I:1-2534000 I:3017001-15058000 II:1-15268000 III:1-2791000 III:11830001-13772000 IV:1-122000 IV:409001-4401000 IV:4474001-5342000 IV:5407001-9002000 IV:9076001-9537000 X:1-4157000
wn106 (wn107)	RIL	7.12	13	I:3035001-10529000 II:3207001-15268000 III:1-12726000 IV:1-220000 IV:388001-667000 IV:740001-9002000 IV:9068001-10931000 IV:10996001-17482000 V:1-2454000 V:2564001-4777000 V:4842001-8196000 X:4898001-13112000 X:13173001-17707000
wn109	RIL	5.86	11	I:1-1616000 I:2624001-11991000 II:12131001-15268000 III:1-3313000 IV:524001-17482000 V:1-2453000 V:2557001-8797000 V:8904001-20913000 X:1-12653000 X:12744001-13101000 X:13176001-17682000
wn111 (wn110)	RIL	7.19	8	I:2890001-15059000 II:1-12305000 IV:13204001-16284000 IV:16345001-17482000 V:1-2486000 V:2557001-19157000 X:1-13112000 X:13176001-17693000
wn113	RIL	4.31	6	I:3242001-15059000 II:13914001-15268000 III:11105001-13772000 V:1-2454000 V:2564001-20913000 X:10346001-13973000

wn116	RIL	4.93	6	II:4584001-15268000 III:8742001-13772000 IV:1-168000 IV:524001-686000 IV:776001-2212000 V:18909001-19702000
wn118 (wn048)	RIL	4.94	8	IV:1-220000 IV:284001-461000 IV:524001-667000 IV:740001-1295000 V:1-2486000 V:2557001-16438000 X:1-1656000 X:12653001-17705000
wn124	RIL	5.36	12	I:2870001-15059000 II:1-711000 II:3903001-12597000 III:1-3910000 IV:388001-570000 IV:642001-1000000 IV:1975001-4401000 IV:4472001-8977000 IV:9041001-11448000 IV:11523001-17482000 V:17975001-20913000 X:14500001-17705000
wn128	RIL	4.77	14	I:1533001-2202000 I:2484001-15059000 II:1-15268000 III:1-4154000 III:4215001-6383000 III:6454001-13767000 IV:1005001-4389000 IV:4472001-8878000 IV:8949001-12633000 IV:12695001-16285000 IV:16346001-17482000 V:1-2486000 V:2557001-8843000 V:8904001-11811000
wn129	RIL	5.64	12	II:11006001-11954000 III:1-4161000 III:4236001-13772000 IV:1-324000 IV:636001-9002000 IV:9068001-10060000 IV:10122001-13313000 V:1-2454000 V:2557001-3921000 X:1-12140000 X:12211001-13112000 X:13173001-17684000
wn134	RIL	6.46	14	I:1-1167000 I:3023001-15059000 III:1-4150000 III:4220001-6397000 III:6465001-13772000 IV:1-122000 IV:620001-9002000 IV:9076001-10810000 IV:10896001-11107000 IV:11184001-17482000 V:1-2454000 V:2527001-8834000 V:8898001-14311000 X:4797001-8487000

wn135	RIL	6.61	16	I:1-2174000 I:2770001-15059000 II:1-15268000 III:2523001-8804000 III:8866001-13087000 IV:69001-220000 IV:388001-524000 IV:620001-2524000 IV:4253001-10357000 IV:10423001-10769000 IV:10835001-10931000 IV:10996001-17482000 V:4413001-8851000 V:8921001-11398000 X:1-8654000 X:13956001-15391000
wn140	RIL	6.29	10	I:1-15059000 II:1-5038000 IV:1-220000 IV:388001-495000 IV:556001-9002000 IV:9076001-16285000 IV:16346001-17482000 V:1-1028000 X:12763001-13112000 X:13176001-17701000
wn142	RIL	6.48	12	I:1-12434000 I:12538001-14012000 I:14086001-15059000 II:1-15268000 III:1-1907000 III:10076001-12617000 IV:1-17482000 V:1-2454000 V:2527001-8845000 V:8921001-20913000 X:1-2881000 X:16562001-17705000
wn146	RIL	6.40	7	I:8022001-15058000 II:1-12985000 IV:1-168000 IV:388001-2063000 V:1-2454000 V:2527001-10163000 X:1-12805000
wn152	RIL	7.49	9	I:1-1567000 I:3101001-15059000 II:1-7960000 IV:1-308000 IV:388001-9002000 IV:9068001-10769000 IV:10835001-17482000 X:11398001-13112000 X:13173001-17705000
wn153	RIL	9.46	8	I:933001-2180000 I:3600001-15058000 III:1-13178000 IV:1336001-17482000 V:1-2486000 V:2564001-8851000 V:8921001-20913000 X:1-13079000

wn158	RIL	5.54	10	I:1-3545000 II:1-15268000 III:1-2822000 III:11198001-12398000 IV:1830001-4401000 IV:4472001-8977000 IV:9041001-17482000 V:1-8766000 V:8851001-20913000 X:3837001-17693000
wn161	RIL	6.89	12	I:2502001-15058000 II:13485001-15268000 III:1-4365000 III:4433001-13772000 IV:1-168000 IV:388001-1408000 IV:1812001-11104000 IV:11166001-13534000 V:1-8821000 V:8904001-9701000 V:9762001-12430000 X:14708001-17705000
wn162	RIL	7.71	8	II:1-3506000 III:4275001-6383000 III:6444001-13769000 IV:2140001-10061000 IV:10122001-15446000 V:3926001-20913000 X:8898001-13112000 X:13173001-17701000
wn171	RIL	7.95	9	II:1-1685000 III:1-13772000 V:1-2486000 V:2557001-4726000 V:4792001-8766000 V:8827001-16038000 X:1-13416000 X:14038001-17327000 X:17396001-17693000
wn176	RIL	8.12	9	I:1-2100000 I:3311001-14698000 II:1-2907000 II:5598001-12010000 III:1-13772000 IV:1979001-14785000 IV:15449001-17482000 V:1-8851000 V:8921001-20913000
wn177	RIL	7.88	5	II:2647001-15268000 III:1-13772000 IV:2629001-15993000 X:2466001-13101000 X:13221001-17707000
wn185	RIL	5.59	7	I:1-1545000 II:12355001-15268000 III:1-1188000 III:2047001-11218000 V:1166001-2454000 V:2557001-8845000 V:8921001-20913000

wn186	RIL	8.54	8	II:1-15268000 III:1-13772000 IV:9719001-17482000 V:1-3867000 V:18463001-20913000 X:1-13112000 X:13173001-13658000 X:14395001-17693000
wn190	RIL	3.11	8	II:1320001-8752000 II:13008001-15268000 IV:2543001-10906000 IV:11075001-16285000 IV:16346001-17482000 V:1-2453000 V:2564001-8766000 V:8851001-17547000
wn196 (wn195)	RIL	5.94	8	II:1-15268000 IV:548001-1677000 IV:1940001-2650000 V:1-2453000 V:2527001-8851000 V:8921001-20913000 X:12446001-12831000 X:13491001-17707000
CB4856	CB4856 control	5.71	13	I:1-15059000 II:1-15268000 III:1-13772000 IV:388001-524000 IV:604001-4389000 IV:4472001-10842000 IV:10909001-17482000 V:1-2486000 V:2557001-8766000 V:8851001-20913000 X:1-9811000 X:9894001-13112000 X:13173001-17705000

Table S3 Coordinates of highly divergent regions in *C. elegans* N2 and CB4856

C. elegans N2 (WS230)	CB4856
I:1086047-1112098	I:1101806-1127574
I:2285977-2305562	I:2327626-2360993
I:2734559-2744454	I:2798869-2809018
I:12476597-12500146	I:12627240-12645122
II:988416-1024139	II:1031447-1074741
II:1470218-1516517	II:1530810-1584979
II:1544781-1579478	II:1613411-1636156
II:1589069-1610149	II:1645699-1671813
II:1686985-1804605	II:1748738-1945322
II:1862640-1887001	II:1993695-2064046
II:1950748-1990258	II:2132999-2184616
II:2044915-2139275	II:2237615-2332746
II:2184771-2201617	II:2388444-2405596
II:2974336-3020963	II:3197502-3250769
II:3097260-3125113	II:3331072-3375129
II:3407722-3456546	II:3672222-3697801
II:3478700-3490495	II:3720082-3730000
III:1-162260	III:1-189633
III:884590-952990	III:931824-994823
III:1302565-1364095	III:1345352-1417238
III:12170290-12200078	III:12350622-12370948
IV:2502463-2518228	IV:2557585-2601617
IV:2609665-2633042	IV:2693368-2721980
IV:3839868-3874858	IV:3938677-3972117
IV:6109360-6140581	IV:6268434-6301517
IV:16036516-16058260	IV:16285400-16352670
V:518047-594265	V:519843-589301
V:687214-723559	V:681654-720089
V:2232729-2264473	V:2259327-2276075
V:2273335-2293636	V:2284931-2303184
V:2393882-2410540	V:2407223-2431542
V:2445201-2455014	V:2580971-2592469
V:2665323-2683253	V:2802647-2821695
V:3133005-3142615	V:3277718-3286975
V:3175100-3211048	V:3319347-3445836
V:3621005-3631859	V:3859108-3870050
V:3667594-3681994	V:3905495-3920741
V:3698216-3715816	V:3939565-3959664
V:7017957-7064447	V:7318802-7360355
V:7294233-7367020	V:7592038-7668476

V:12116567-12130512	V:12433464-12449980
V:15394389-15439019	V:15722924-15764693
V:15580092-15674206	V:15889637-15980152
V:15865905-15917533	V:16175096-16243084
V:16134646-16181460	V:16463162-16501406
V:16270370-16280562	V:16595350-16607029
V:16545600-16553689	V:16848550-16856085
V:16605149-16633418	V:16907288-16963843
V:16870092-16958775	V:17225619-17367259
V:17256246-17270779	V:17773246-17787294
V:17285255-17296361	V:17801498-17834237
V:17645339-17710126	V:18193641-18260001
V:18085929-18094141	V:18675982-18691917
V:18734280-18876232	V:19366735-19527624
V:19304333-19313907	V:19986300-19996675
V:19442863-19465521	V:20127985-20173435
V:19492868-19522263	V:20203259-20235306
V:19763911-19775102	V:20482152-20493438
V:19819367-19884716	V:20540845-20625910
X:1507699-1534961	X:1534338-1561578
X:14046536-14057565	X:14217202-14228221

Table S4 Tc1 copies in the N2 and CB4856 genomes

Retained					
copies	Tc1 start	Tc1 end	N2 coordinates	CB4856 coordinates	Clone
1	1	1610	chrI:10528229-10529839	I:10448092-10449702	F59C6
2	1	1612	chrI:12481317-12482892	I:12329190-12330765	T27F6
3	1	1610	chrI:14021864-14023474	I:13858652-13860198	ZC334
4	1	1610	chrII:6571191-6572801	II:6301769-6303379	F18C5
5	1	1472	chrII:13307543-13309014	II:12983176-12984647	Y48C3A
6	1	1610	chrIII:2119183-2120793	III:2057190-2058796	Y48G9A
7	1	1610	chrIII:12865065-12866675	III:12688772-12690382	Y37D8A
8	1	1610	chrIV:9685246-9686856	IV:9514675-9516285	ZK1251
9	1	1610	chrIV:13223694-13225304	IV:13030790-13032400	C48D1
10	1	1610	chrIV:15484067-15485677	IV:15253992-15255602	Y73F8A
11	90	1610	chrV:1834327-1835848	V:1818396-1819853	T10B5
12	1	1610	chrV:3989832-3991442	V:3745971-3747581	B0213/K09D9
13	1	182	chrV:9890310-9890491	V:9575212-9575393	C50H2
	876	1610	chrV:9890486-9891220	V:9831139-9831873	
14	1	1610	chrV:10187631-10189240	V:9872535-9874144	ZK856
15	1	112	chrV:11288609-11288720	V:10973599-10973710	C03E10
	68	1619	chrV:11288742-11290273	V:11229563-11231094	
16	1	1610	chrV:16312007-16313617	V:15984106-15985716	C31A11
17	1	1610	chrV:18442316-18443926	V:17859623-17861233	Y51A2C/ZK228
18	1	1610	chrX:7017195-7018805	X:6882401-6884011	R173
19	1	1610	chrX:11311136-11312746	X:11149404-11151014	F08G12

Deleted					
Full length	Tc1 start	Tc1 end	N2 coordinates		Clone
1	1	860	chrII:4168682-4169542		R03H10
	1535	1610	chrII:4169535-4169610		
2	1	1610	chrII:784390-786001		Y39F10A/F46F5
3	1	1610	chrII:894218-895828		T07D3
4	1	1610	chrII:1935486-1937097		ZK250
5	1	1610	chrII:7516627-7518237		C28F5
6	1	1610	chrII:12520686-12522296		T21B4
7	1	1610	chrII:12750742-12752352		Y46G5A
8	1	1610	chrIV:2485118-2486727		Y69A2AR
9	1	1610	chrV:1836070-1837680		Y32G9B/T10B5[a]
10	1	1610	chrV:3251322-3252932		T28A11[a]
11	1	1610	chrV:3590233-3591844		T22F3[a]
12	1	1610	chrV:15907121-15908731		F35E8
13	1	1610	chrV:17805773-17807383		Y94A7B

Inserted				
One end only				
orientation	Tc1 start	Tc1 end	CB4856 coordinates	
-	1525	1610	I:1645408-1645496	1
+	1487	1610	II:3089263-3089386	2
+	1	73	III:13494864-13494936	3
-	1557	1610	V:12058803-12058856	4
+	1	83	V:17660363-17660445	5

Inserted -								
Both ends								
orientation	Tc1 start	Tc1 end	CB4856 coordinates	orientation	Tc1 start	Tc1 end	CB4856 coordinates	
+	1	100	I:2635117-2635212	+	1539	1610	I:2635203-2635274	1
+	1	82	I:12276809-12276890	+	1534	1610	I:12276886-12276962	2
+	1	68	II:1929985-1930051	+	1534	1610	II:1930048-1930124	3
-	1530	1610	II:3087013-3087093	-	1	79	II:3087086-3087164	4
-	1	81	II:7549244-7549324	+	1533	1610	II:7549172-7549247	5
+	1	77	II:12569787-12569863	+	1534	1610	II:12569787-12569863	6
-	1535	1610	II:13612906-13612981	-	1	72	II:13612982-13613053	7
+	1	100	II:14472906-14473004	+	1534	1610	II:14472990-14473066	8
-	1533	1610	V:2775116-2775193	-	1	78	V:2775189-2775266	9
-	1525	1610	V:7189046-7189131	-	1	80	V:7189132-7189211	10
-	1542	1610	V:12390311-12390379	-	1	72	V:12390380-12390451	11
+	1	84	V:16571603-16571686	+	1546	1610	V:16571687-16571751	12

[a]Present in DOLGIN et al. 2008

Table S5 SNVs excluding divergent regions

Chromosome	Bases in ARMS	SNVs on ARMS	Fraction SNVs on ARMS	Bases in CENTERS	SNVs on CENTERS	Fraction SNVs on CENTERS
I	6590733	23628	0.359	8163409	8624	0.106
II	6261468	29100	0.465	7941885	8843	0.111
III	6298078	19547	0.31	6930366	5429	0.078
IV	8102292	23349	0.288	8865871	7192	0.081
V	10021279	42785	0.427	8803113	8607	0.098
X	10046187	12058	0.12	7427535	8076	0.109

Table S6 Previously studied loci for *C. elegans* N2/CB4856

Locus	Study	Genotype	Identified same change found in cited reference
rtw-5	(HODGKIN AND DONIACH 1997)	In N2 *rtw-5* carries a variant G to A change at position 46… CB4856 has the ancestral genotype	yes
npr-1	(DE BONO AND BARGMANN 1998)	One isoform, NPR-1 215F, is found exclusively in social strains, while the other isoform, NPR-1 215V, is found exclusively in solitary strains. An NPR-1 215V transgene can induce solitary feeding behavior in a wild social strain." "Within the *npr-1* locus, the two social strains CB4856 and CB4932 both had an insertion of approximately 50bp in intron 1…"	yes
ppw-1	(TIJSTERMAN *et al.* 2002)	"Sequencing this locus identified multiple polymorphisms between the Hawaiian strain and N2. Besides some base alterations that result in the alterations of the amino acid sequence (including two amino acid changes in protein domains that are highly conserved), the Hawaiian allele carries a single base deletion resulting in an early stop codon that is suggestive of a null allele." ppw-1	4 NS and 1 SS change; one frameshift and a inframe indel
	(ELVIN *et al.* 2011)	We first sequenced the *ppw-1* locus for 31 RILs and then analyzed whether the genotype can predict the phenotype…"Polymorphisms in ppw-1 appear to explain the loss of germline RNAi sensitivity in Hawaii"	
	(POLLARD AND ROCKMAN 2013)	"we distinguished N2 and CB alleles of *ppw-1* by sequencing through chrI, which is the position of a 1bp deletion… a 5kb fragment was amplified necessary to avoid amplifying the paralog *sago-2*. "	yes
gst-38	(DENVER *et al.* 2003)	"the divergent F35E8 locus on chromosome V spanned the entire F35E8.8 gene, composed of two exon and one intron. All of the variable sites detected at this locus were base substitutions… the majority of substitutions detected at this locus were in exon sequences. Among the 30 base substitutions observed in F35E8.8 exon sequence, eight were replacements and 22 were silent. No base substitutions were observed that resulted in premature stop codons for the F35E8.8 gene."	yes
	(MAYDAN *et al.* 2007)	We also examined a gene, *gst-38*, that has been sequenced from the Hawaiian strain and is known to have several SNPs relative to Bristol [Denver 2003]. Probe targets in the Hawaiian genome contain 0-3 SNPs each, which resulted in a significantly negative log2 ratio in that region of the genome, but not of sufficient amplitude to pass our conservative criteria for identifying deletions."	NA

tra-3	(KAMMENGA *et al.* 2007)	The natural variation in body size response to temperature between CB and N2 was caused by a single mutation F96L in a calpain-like protease TRA-3 encoded by *tra-3*….One SNP was found within the coding region where phenylanaline-96 in N2 was mutated into leucine-96 in CB	no
C47G7.1, D1065.3	(MAYDAN *et al.* 2007)	"We identified a 2943bp deletion on chrV in the Hawaiian strain CB4856 that affects two adjacent genes, C49G7.1 and D1065.3. Both are uncharacterized genes… we designed primers flanking the deletion, amplified the affected region using PCR, and sequenced the region to determine the deletion breakpoints."	yes
plg-1	(PALOPOLI *et al.* 2008)	Sequence analysis revealed that the [Cer1] retrotransposon and its long terminal repeats (LTRs) interrupts a novel, unannotated protein-coding gene whose predicted product has similarity to canonical mucins… the protein is predicted to contain proline, threonine, and serine-rich (PTS) repeats." ["We show that the plugging polymorphism results from the insertion of a retrotransposon into an exon of a novel mucin-like gene, plg-1, whose product is a major structural component of the copulatory plug."]	8.8 kb deletion in CB4856 relative to N2
peel-1 / zeel-1	(SEIDEL *et al.* 2008)	"The interval to which *zeel-1* and *peel-1* map contains a region of dramatically elevated sequence divergence between the Bristol and Hawaii haplotypes. This region spans 33kbs of Bristol sequence and includes four full genes and part of a fifth. The Hawaii haplotype contains a 19-kb deletion spanning the gene *Y39G10AR.5*. Divergence within coding segments of the remaining genes averages 5%, which is 50 times higher than previous genome-wide estimates of pairwise divergence from both coding and noncoding sequence. Noncoding segments in this region are largely unalignable and contain many insertions and deletions, mainly composed of repetitive elements. The left boundary of the divergent interval is abrupt and is marked by a 1-kb insertion in Hawaii. Genomic divergence within the 13kb immediately outside the insertion is 0.1%. The right boundary is less abrupt, with divergence falling gradually to 0.7% across 4kb."	yes
tyra-3	(BENDESKY *et al.* 2011)	"To identify polymorphisms between N2 and HW alleles of *tyra-3*, we sequenced ~19kb surrounding the *tyra-3* locus in HW. There were 34 differences between HW and the N2 consensus genomic sequence: 33 noncoding changes and a single coding difference that changed a glutamate in the *tyra-3b* isoform to glycine."	yes

		"*tyra-3* transgenes with the N2 noncoding sequence were significantly more potent than comparable transgenes with the HW sequence…excluding the coding polymorphism and localizing a functional difference between N2 and HW *tyra-3* genes to a 4.9kb region that harbors 5 noncoding SNPs, one 1bp insertion, and a 184bp deletion in HW… results indicate that the 184bp deletion represents at least part of the functional difference between N2 and HW *tyra-3* alleles. "	
abts-3/exp-1	(BENDESKY *et al.* 2011)	Sequencing this region uncovered 11 polymorphisms between HW and N2: five noncoding SNPs, two coding SNPs, one 1bp deletion, one 1bp insertion, a 3bp insertion, and a 23bp deletion." (abts-3) "a second gene close to the II-QTL is *exp-1*… the stop codon of *exp-1* is 2.2kb away from the QTL…. we suggest that a noncoding variation 3' of the *exp-1* transcript, within the *abts-3* gene, modifies aggregation and bordering behavior by affecting the activity of *exp-1*. "	yes, plus additional SNVs and indels
glc-1	(GHOSH *et al.* 2012)	To identify the functional polymorphism(s), we sequenced the N2 and CB alleles of *glc-1*. Relative to N2, CB had 77 SNPs in the coding region, 32 of which resulted in AA changes, as well as a 4-AA deletion in exon 2. Despite the multiple coding polymorphisms, the predicted secondary structure and membrane topology of GLC-1 from N2 and CB were similar… "The *glc-1* region exhibited high sequence divergence between N2 and CB, with 178 SNPs in 5kb, a polymorphism rate ~30 times higher than the average of 1SNP/840bp (Wicks). Sequences of five other gluatmate-gated Cl channel subunits differed very little between N2 and CB."	253 SNVs
tac-1	(TARAILO-GRAOVAC AND CHEN 2012)	"In CB we identified a number of SNVs that affect *tac-1*. In particular, one variation affected the second exon of *tac-1* causing a C94W change in the AA sequence."	yes
glb-5	(PERSSON *et al.* 2009) (MCGRATH *et al.* 2009)	This interval contained two genes, one of which, *glb-5*, had similarity to globins. Comparison of N2 and CB4856 sequences in the 8kb revealed 11 polymorphisms, all in *glb-5*. Ten SNPs altered introns, the remaining polymorphism partially duplicated glb-5 in N2 but not CB4856. "The genomic sequence of *glb-5* contained a 765bp duplication/insertion in N2 compared to HW. *glb-5* cDNA analysis demonstrated that the DNA polymorphism resulted in substantially different mRNAs and predicted GLB-5 proteins in N2 and CB. The duplicated exon in N2 led to an in-frame stop codon in the *glb-5* cDNA, resulting in a truncation of the last 179 AAs of the protein compared to HW, and the inclusion of 40 different residues."	9 SNPs in introns, 3 indels including one the deletion in hawaiian corresponding to the partial duplication of glb-5 in N2 yes

Supplementary References

Bendesky, A., M. Tsunozaki, M. V. Rockman, L. Kruglyak and C. I. Bargmann, 2011 Catecholamine receptor polymorphisms affect decision-making in C. elegans. Nature 472**:** 313-318.

de Bono, M., and C. I. Bargmann, 1998 Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in C. elegans. Cell 94**:** 679-689.

Chu, T. C., C. H. Lu, T. Liu, G. C. Lee, W. H. Li *et al.*, 2013 Assembler for de novo assembly of large genomes. Proc Natl Acad Sci U S A 110**:** E3417-3424.

Denver, D. R., K. Morris and W. K. Thomas, 2003 Phylogenetics in Caenorhabditis elegans: an analysis of divergence and outcrossing. Mol Biol Evol 20**:** 393-400.

Dolgin, E.S., B. Charlesworth, A.D. Cutter, 2008 Population frequencies of transposable elements in selfing and outcrossing Caenorhabditis nematodes. Genet Res 90:317-329.

Doroszuk, A., L. B. Snoek, E. Fradin, J. Riksen and J. Kammenga, 2009 A genome-wide library of CB4856/N2 introgression lines of Caenorhabditis elegans. Nucleic Acids Res 37**:** e110.

Elvin, M., L. B. Snoek, M. Frejno, U. Klemstein, J. E. Kammenga *et al.*, 2011 A fitness assay for comparing RNAi effects across multiple C. elegans genotypes. BMC Genomics 12**:** 510.

Ghosh, R., E. C. Andersen, J. A. Shapiro, J. P. Gerke and L. Kruglyak, 2012 Natural variation in a chloride channel subunit confers avermectin resistance in C. elegans. Science 335**:** 574-578.

Grishkevich, V., S. Ben-Elazar, T. Hashimshony, D. H. Schott, C. P. Hunter *et al.*, 2012 A genomic bias for genotype-environment interactions in C. elegans. Mol Syst Biol 8**:** 587.

Hodgkin, J., and T. Doniach, 1997 Natural variation and copulatory plug formation in Caenorhabditis elegans. Genetics 146**:** 149-164.

Hunt, M., T. Kikuchi, M. Sanders, C. Newbold, M. Berriman *et al.*, 2013 REAPR: a universal tool for genome assembly evaluation. Genome Biol 14**:** R47.

Kammenga, J. E., A. Doroszuk, J. A. Riksen, E. Hazendonk, L. Spiridon *et al.*, 2007 A Caenorhabditis elegans wild type defies the temperature-size rule owing to a single nucleotide polymorphism in tra-3. PLoS Genet 3**:** e34.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25**:** 2078-2079.

Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.*, 2006 Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet 2**:** e222.

Maydan, J. S., S. Flibotte, M. L. Edgley, J. Lau, R. R. Selzer *et al.*, 2007 Efficient high-resolution deletion discovery in Caenorhabditis elegans by array comparative genomic hybridization. Genome Res 17**:** 337-347.

McGrath, P. T., M. V. Rockman, M. Zimmer, H. Jang, E. Z. Macosko *et al.*, 2009 Quantitative mapping of a digenic behavioral trait implicates globin variation in C. elegans sensory behaviors. Neuron 61**:** 692-699.

Palopoli, M. F., M. V. Rockman, A. TinMaung, C. Ramsay, S. Curwen *et al.*, 2008 Molecular basis of the copulatory plug polymorphism in Caenorhabditis elegans. Nature 454**:** 1019-1022.

Persson, A., E. Gross, P. Laurent, K. E. Busch, H. Bretes *et al.*, 2009 Natural variation in a neural globin tunes oxygen sensing in wild Caenorhabditis elegans. Nature 458**:** 1030-1033.

Pollard, D. A., and M. V. Rockman, 2013 Resistance to germline RNA interference in a Caenorhabditis elegans wild isolate exhibits complexity and nonadditivity. G3 (Bethesda) 3**:** 941-947.

Seidel, H. S., M. V. Rockman and L. Kruglyak, 2008 Widespread genetic incompatibility in C. elegans maintained by balancing selection. Science 319**:** 589-594.

Smit, A. F. A., R. Hubley and P. Green, 1996-2010 RepeatMasker Open-3.0.

Tarailo-Graovac, M., and N. Chen, 2012 Mos1-mediated transgenesis to probe consequences of single gene mutations in variation-rich isolates of Caenorhabditis elegans. PLoS One 7**:** e48762.

Thompson, O., M. Edgley, P. Strasbourger, S. Flibotte, B. Ewing *et al.*, 2013 The million mutation project: a new approach to genetics in Caenorhabditis elegans. Genome Res 23**:** 1749-1762.

Tijsterman, M., J. Pothof and R. H. Plasterk, 2002 Frequent germline mutations and somatic repeat instability in DNA mismatch-repair-deficient Caenorhabditis elegans. Genetics 161**:** 651-660.

Wicks, S. R., R. T. Yeh, W. R. Gish, R. H. Waterston and R. H. Plasterk, 2001 Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. Nat Genet 28**:** 160-164.